Units of Analysis for the Legal Domain: A Legal Document Ontology

William Mandrick[1,†,*] and Olivia Hobai[2,†]

[1] CUBRC, 4455 Genesee Street, Buffalo, New York 14225, United States of America
[2] University at Buffalo, 12 Capen Hall, Buffalo, New York 14260-1660, United States of America

Abstract

Machine interpretation (MI) within the legal domain requires a consistently structured ontology that applies across a corpus of interrelated documents—e.g., the U.S. Code or U.S. Code of Federal Regulations. The proposed Legal Document Ontology (LDO) clearly distinguishes between structural and semantic (i.e. meaningful) entities as the two primary units of analysis. It also maintains a set of standardized relationships (object properties) for human and computer reasoning with both structural and semantic entities. The structural entities of a legal document, such as a clause or subsection, relate to each other as they compose the document itself. In contrast, semantic content entities relate to other entities by being about them. In what follows we will use elements of the Document Components Ontology (DoCo) and the Common Core Ontologies (CCO) to justify and establish the first version of the LDO.

Keywords

Legal Documents, Machine Interpretation, Applied Ontology, Information Structure Entity, Information Content Entity

1. Introduction

We intend to contribute to the ontology of law for machine interpretation (MI), one that applies to both individual legal documents and entire corpora such as the U.S. Code or Code of Federal Regulations. We advocate an applied ontology that formally represents the entities and relationships within a legal text. Moreover, this Legal Document Ontology (LDO) exists to enhance the automated reasoning of legal documents with structured, graph-computable knowledge of the legal domain.

The motivation for this effort in applied ontology comes from the ongoing debates on how best to interpret the word of law. Legal experts must wade through linguistic ambiguities and incompatible theories of jurisprudence from within these debates. This is where Machine Interpretation offers powerful benefits beyond basic keyword search or text matching. MI is the computational process by which the semantic content of documents is automatically extracted, represented, and reasoned over, using techniques from natural language processing [1], knowledge representation, and logic-based inference [2]. In document analytics [3], machine interpretation lets systems go beyond surface-level text processing to derive structured meaning [4], identify implicit relationships, detect intent or anomalies, and support automated or human-in-the-loop decision-making[1]. Machine interpretation would thus greatly benefit legal practitioners in determining the scope of a statute or the relevant legislation that supports it. In this way the LDO models the structure and content of legal documents to facilitate computer-assisted reasoning over large amounts of data.

The LDO is compliant with the Basic Formal Ontology (BFO) [5, 6] and the Common Core Ontologies (CCO) [7, 8]. Both provide a general classification scheme from which to build an ontology of law. This proposal centers the classification of information as a generically dependent continuant:

Proceedings of the Joint Ontology Workshops (JOWO) - Episode XI: The Sicilian Summer under the Etna, co-located with the 15th International Conference on Formal Ontology in Information Systems (FOIS 2025), September 8–9, 2025, Catania, Italy

* Corresponding author.
✉william.mandrick@cubrc.org (W. Mandrick); oihobai@buffalo.edu (O. Hobai)
† These authors contributed equally.

1 Definition synthesized by the authors, drawing on foundational work in natural language processing [1], ontology engineering [2], semantic document analytics [3], and linked data/knowledge graphs [4].

(BFO) generically dependent continuant =def. an entity that exists in virtue of the fact that there is at least one of what may be multiple copies which is the content or the pattern that multiple copies would share.

Indeed, every instance of a law can exist in the writing of legal textbooks, in online databases, or in the memory of legal officials. These various resources carry one and the same legal content or pattern. Furthermore, what differentiates one law from another *is* their content or pattern. Here we define that content or pattern as a subtype of the Process Regulation class in the CCO, which itself is a subtype of the Information Content Entity. The definitions for those are as follows:

(CCO) Process Regulation =def. A Directive Information Content Entity that prescribes a Process as required, prohibited, or permitted, and is the output of a Process which realizes some Authority Role.

(CCO) Information Content Entity =def. A Generically Dependent Continuant that generically depends on some Information Bearing Entity and stands in relation of aboutness to some Entity.

The GDC is the foundation for more specifically relevant subclasses in the CCO — namely the Information Content Entity (ICE). The LDO extends the subclasses of the ICE in order to represent the semantic *units of analysis* of a legal document; those being the ICEs about agents and processes.

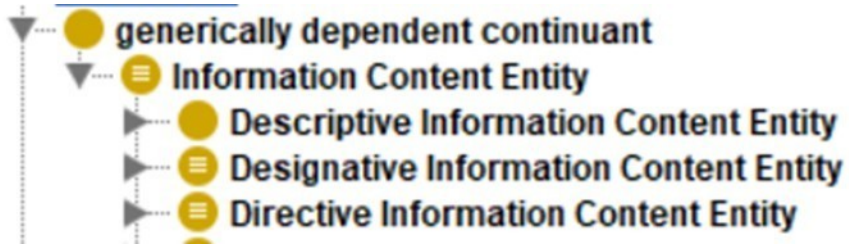

Figure 1: The BFO provides the class 'generically dependent continuant' as the foundation for a taxonomy of Descriptive, Designative, and Directive Information Content Entities.

In contrast, the CCO class of Document Field represents the structural elements of a legal document: the *parts*, *sections*, *sub-sections*, etc. Legal instances of a Document Field are *carriers of* some semantic content: a definition, authorization, prohibition, etc. *of* some process. For our purposes, the Information Bearing Entity that is a Document Field is only of indirect interest, as we hold that one can analyze the information content of a law from the formatting or structure that carries it. It is this analysis that can put the same law in the mind of a lawyer, a legal textbook, and an online database.

2. Justification

The creation and enforcement of a law brings certain entities into existence: the duration of a jail sentence, for instance, or the imposition of a civil status such as *felon* or *married*. In this sense, laws are ontological statements. Using BFO and CCO terminology, a law about citizenship *authorizes* some citizen *role* of a person. The law is then relevant to determining whether that citizen role endures over time, the processes one can participate in, etc. Citizenship is a *legal status* that directly depends on the authority of a government to legislate over individual citizens. J.L. Austin's speech act theory provides a philosophical grounding for understanding law as action through language. Laws are not just propositions or texts; they are declarations that change social and legal reality when performed correctly. This has deeply influenced legal theory, especially in jurisprudence, legal positivism, and constitutional theory. He states, "To say something is to do something, or in

saying something we do something." [9] These performative statements are 'directive information content entities' in the CCO.

The ontological distinction between structural and semantic elements has implications for both legal analysis and digital publishing. Whereas legal analysis concerns itself with the creation and application of law qua directive ICE, digital publishing transforms law from a static, paper-based institution into a dynamic, interconnected digital ecosystem. It enhances access and efficiency, but also demands new standards for authenticity, interpretation, and preservation. Ultimately, it is reshaping how legal systems function in the digital age—making the law not just readable by people, but increasingly interpretable by machines.

Legislating over individual citizens, in turn, directly involves also the relationships *between* laws, perhaps at the level of titles, sections, and subsections. These forms of structure are but a *vehicle* for the content of a legal statement. The order of conditions is less important than the conditions themselves. The complex nature of the legal domain means a single Legal Information Bearing Entity (IBE) can end up operating in a broad range of applications. For example, a footnote in a ruling that the government had the right to strict scrutiny regarding fillers in milk (*United States v. Carolene Food Products*) had downstream effects in justifying strict scrutiny regarding civil rights discrimination [10]. The relevance of footnote 4 to later cases exists because of the ontological claim it made about the Authority Role of the Government. It is the *content* that develops new legal arguments and applications, rather than its placement in the structure of the Supreme Court opinion. The structure that the footnote occupies merely exists for human comprehension.

The Common Core Ontologies (CCO), Information Artifact Ontology (IAO) [11] and Document Components Ontology (DoCO) [12], all acknowledge this structure/content distinction in different ways. One can contrast the LDO proposed here with the taxonomical structures of the IAO as well as the DoCO. The latter classification schemes primarily work in terms of classifying information structure. We hold that the Common Core-compliant alternative proposed here complements both the IAO and the DoCo; LDO takes the salient parts of all three ontologies to result in a maximally expressive account of the legal domain.

The structural units of analysis in the referenced ontologies are depicted below:

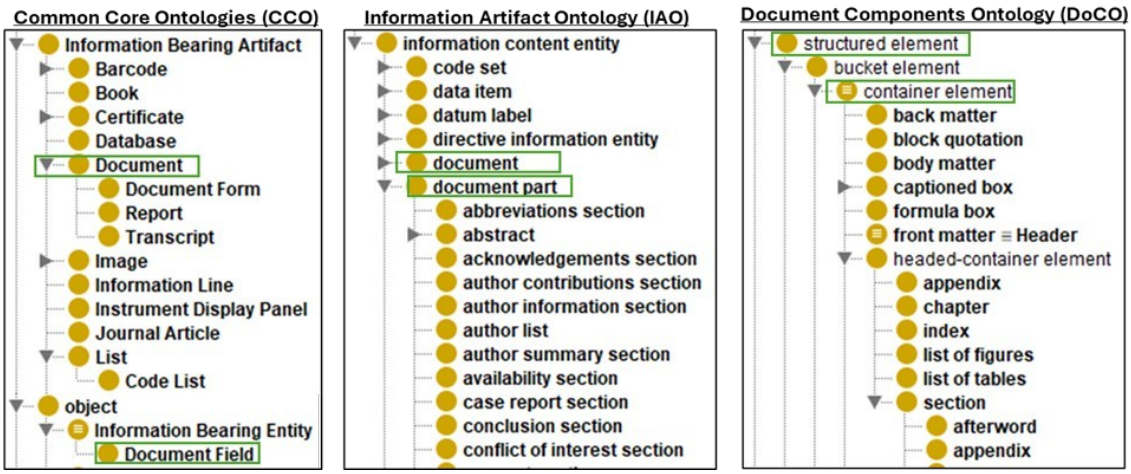

Figure 2: The CCO, IAO, and DoCo all make distinctions between the physical and relational structure of a document— marked by the green boxes around CCO 'Document' and 'Document Field', IAO 'document' and 'document part', and DoCO 'structured element' and 'container element'. The ontologies here present three different ways of taxonomically arranging the units of document structure.

The IAO and DoCO then appear to combine structure and content into classes such as *case report section* (IAO) and *captioned box* (DoCO). The DoCO distinguishes between a *structured element* defined as "an element that can contain other elements" and a *discourse element* defined as "an element of a document that carries out a rhetorical function." We take the discourse element as an analogous class to the CCO Information Content Entity. Rhetoric exists in formal arguments, and

every argument is necessarily about some object. However, problems arise upon looking at the discourse element subclasses: namely, such terms as *footnote*, *front matter*, and *label*. These three classes, alongside other discourse elements, frequent legal documents sufficiently for us to advocate their use in an ontology of the legal domain. The problem is that the rhetorical function of the discourse element is mistakenly combined with the structural entities. The class *label* is defined as follows:

> (DoCO) label =def. A block containing text, that may include a number (e.g., "Chapter Three", "3.2", "Figure 1", "Table"), used to identify an item within the document, for example a chapter, a figure, a section or a table.

Indeed, labels identify items within documents; the "block containing text", however, is a separate entity. The very same label could exist in multiple different blocks across multiple documents. The defining feature of a label is, as the DoCO agrees, its *use*. The block in the document *containing* said label is a CCO Document Field; the label is a CCO Designative Information Content Entity.

The class *footnote* is defined with a similar conflation:

> (DoCO) footnote =def. A structure within a sentence that permits the author to make a comment or to cite another publication in support of the text, or both. A footnote is normally flagged by a superscript number immediately following that portion of the text to which it relates. For convenience of reading, the text of the footnote is usually printed at the bottom of the page or at the end of a text.

The analysis of content from structure in a footnote calls for an ontological distinction between (1) a structure within a sentence and (2) the comment that sentence carries. It is this latter use that matters for jurisprudence. We argue the meaning of a legal footnote is independent of the structure; footnote 4 could be at the very beginning or the very end of the Supreme Court Opinion. The "rhetorical function" of the comment transcends the position it has in the text. The content of the footnote bears a defining relation of aboutness to the authority of the U.S. government, not to the other structural parts.

The units of analysis for the LDO read meaning away from structure, rather than uniting the two. Information structures themselves are void of meaning. While the DoCo and the IAO prove helpful in *parsing* a legal document, *interpreting* that document can only be in terms of the information content. The LDO takes the relevant *structural* units of analysis that the DoCo describes as discourse elements as being the rightful subclasses of CCO Document Field. The rhetorical carried by a particular instance of Document Field is a separate ICE. What results from this is an account of document parthood along both structural and semantic axes, thus allowing for clear parsing and inference.

3. Legal Document Ontology Classes

Each class in the LDO is a formal representation of a category of entities that share common characteristics. Inspired by the CCO, DoCO and IAO, the LDO classes for document structure are as follows:

> (CCO) Document =def. An Information Bearing Artifact (IBA) that is designed to bear some specific Information Content Entity in a series of paragraphs of text or diagrams in the form of physical pieces of paper or an electronic word processor file.
>
> (CCO) Document Field =def. An Information Bearing Entity (IBE) that is a part of some document into which bearers of prescribed information can be written or selected.
>
> (LDO) Legal Document Part =def. A Document Field considered to be an intermediate division that includes a title.

(LDO) Legal Document Chapter =def. A Legal Document Part that is a division or section of a written work of law to organize content.
(LDO) Legal Clause =def. A Legal Document Part that is the carrier of a discrete proposition, obligation, condition, exception, right, or stipulation
(LDO) Legal Document Section =def. A Legal Document Part (structural division) that contains a logically coherent group of provisions, clauses, or statements.

To model the *content of* an IBE we follow the three CCO units of analysis: (1) designation, (2) description, or (3) prescription. Parsing descriptive from prescriptive content in this manner rightfully responds to the issue of ontological ambiguity in legal claims. In *A Treatise of Human Nature*, David Hume [13] observed that many writers move from statements of fact (about the world, human behavior, or society) to moral claims about what people should do, without explaining the justification for that transition. In this way Hume highlights a logical gap between descriptive statements (what *is*) and prescriptive or normative statements (what *ought* to be). Using this analysis, the LDO could allow for easier reference to the subject of a law and the legal status prescribed to it, rather than assuming a univocal meaning to every legal statement.

The first version of the classes in the LDO follow this CCO-compliant taxonomical hierarchy:

(CCO) Designative Information Content Entity =def. An Information Content Entity that consists of a set of symbols that denote some Entity.
(LDO) Title =def. A designative information content entity that is a name or label given to a written work (e.g. a law) to identify or describe it.

In designative instances, ICEs assert the existence of an entity via a title, name, address, or serial numbers. Some examples of Legal Designative Information Content Entities are license plate numbers, social security numbers, or the legal name on one's birth certificate.

Other kinds of ICEs describe, represent, or measure entities:

(CCO) Descriptive Information Content Entity =def. An Information Content Entity that consists of a set of propositions that describe some Entity.
(LDO) Definition =def. A descriptive information content entity that explains the meaning of a word, phrase, concept, or object. It identifies the essential qualities that distinguish the thing being defined from all other things.

One example of a legal description would be that of an act requirement in a criminal statute, the necessary and sufficient conditions under which a crime is committed. Legal definitions cover:

- Person Roles such as "public official" and "immigrant"
- Objects such as "real property" and "firearm"
- Actions such as "assault" and "commercial fishing"
- Events such as "breach of contract" and "natural disaster"
- Measurements such as "business day" and "safe concentration level"

Legal definitions describe who is covered, what is regulated, where it applies, when it applies, how it's measured, and what specific terms mean in context. Legal definitions are the foundation that determines the reach, limits, and enforceability of the law.

Descriptions then provide the foundation for prescriptive claims such as requirements or recommendations— one must first define what is meant by *assault* before prohibiting it. Directive ICEs then prescribe, authorize, prohibit, or require action:

(CCO) Directive Information Content Entity =def. An Information Content Entity that consists of a set of propositions or images (as in the case of a blueprint) that prescribe some Entity.

A law can *prohibit*, *permit*, or *require* some process. A Criminal Law *prohibits* some Criminal Act (e.g. a homicide) and *permits* another Legal System Act (e.g. a criminal conviction). Alongside the premise that every law is a Process Regulation in CCO terms, we put forward the view that a legal prescription is always about some *action*: all acts that obey the law are kinds of processes, as are the acts that break it. For instance, criminal laws about illegal firearms are about *prohibiting some acts of obtaining a firearm*, rather than *the firearms themselves*.

As defined in the Introduction, the first condition of a Process Regulation represents how every law has a prescriptive force on some kind of Process. The second condition represents how that prescription depends on some agent (the Government) and a (Government) Authority Role. The Authority Role of the Government Legislature, a Government Organization that is part of the Government, is what then enforces accountability for those prescriptions. The United States Federal and State Governments, along with their subsidiary agencies, make a given prescription legally binding. Even those legal prescriptions that hold between civilians, such as contracts, are part of state and federal jurisdiction.

Given the motivation above, we classify Law in the Legal Document Ontology as follows:

(LDO) Law =def. A Process Regulation that (1) prescribes some Act that has_agent some Person and (2) is the output of a Legal System Act that has_agent some Government Legislature and realizes the Authority Role of that Government.

Users of the LDO can further restrict the Act conditions of (1) to represent a specific area of law. What follows is an application of the definition to criminal law only:

(LDO) Criminal Law =def. A Process Regulation that (1) prohibits some Criminal Act that has_agent some Person and (2) is the output of a Legal System Act that has_agent some Government Legislature and realizes the Authority Role of that Government Legislature.

Additionally, some laws hold federally, but not in every State. By complement, some laws only hold at the State level and are outside federal jurisdiction. Sometimes laws will differ between counties in one and the same State. This means that laws at different levels of jurisdiction simply redefine condition (2), as shown:

State Law =def. A Process Regulation that (1) prescribes some Act that has_agent some Person and (2) is the output of a Legal System Act that has_agent some State Government and realizes the Authority Role of that State Government

By merely reclassing the relevant Agent and Authority Role in the Legal System Act, one can represent legal prescriptions at various levels of government. Instead of housing all laws under the most generic class, differences in subject matter and scope can map onto data sets for targeted reasoning.

4. Object Properties for the Legal Document Ontology

Machine readable object properties (relationships) have a critical role in enhancing the machine interpretability of a document. Object properties standardize relations, allowing a system to align content from different sources. They facilitate complex queries that go beyond keyword search by leveraging the structure of knowledge obtained about multiple entities.

The object properties we are proposing for the LDO facilitate human and computer reasoning regarding both the structure and content of a legal document. For example, the object property 'has member part' allows humans and computers to reason about the composition of a legal document. This facilitates queries to identify all the documents in a corpus that contain a certain 'document field'. In contrast, the object property 'prohibits' is a relationship between a legal prohibition (a law) and some activity. The proposed relationships in Table 1 facilitate queries that start with "what laws or regulations prohibit (or authorize or allow)". Object properties such as 'has member part' facilitate reasoning regarding document structure whereas object properties such as 'requires' and 'is prohibited by' facilitate reasoning regarding relations of aboutness.

Table 1
Object properties for the Legal Document Ontology

Object Properties	Inverse Object Properties
Document *has member part* Section	Section *member part of* Document
Document Field *is carrier of* Law	Law *generically depends on* Document Field
Law *is about* Activity	Activity *is subject of* Law
Description *describes* Entity	Entity *described by* Description
Title *designates* Law	Law *designated by* Title
Process Regulation *prescribes* Activity	Activity *prescribed by* Process Regulation
Law *permits* Activity	Activity *permitted by* Law
Law *prohibits* Activity	Activity *prohibited by* Law
Law *requires* Activity	Activity *required by* Law

5. Applications

The United States code is the official compilation of the general and permanent laws of the United States organized by subject matter. Enacted by Congress, the U.S. code serves as the foundation for federal statutory law and is used by legal professionals, lawmakers, and the public to understand the structure and application of the law. In line with the ontology development principles we advocate here, the structure of the Code must undergo frequent standardization measures to ensure clarity, consistency, and hierarchical organization. These standards enable users to locate and interpret statutes efficiently.

The U.S. Code's structural elements are scaffolding for its semantic content. Typically, titles and chapters organize the law into coherent units of analysis, while sections and subsections contain specific legal definitions and prescriptions.

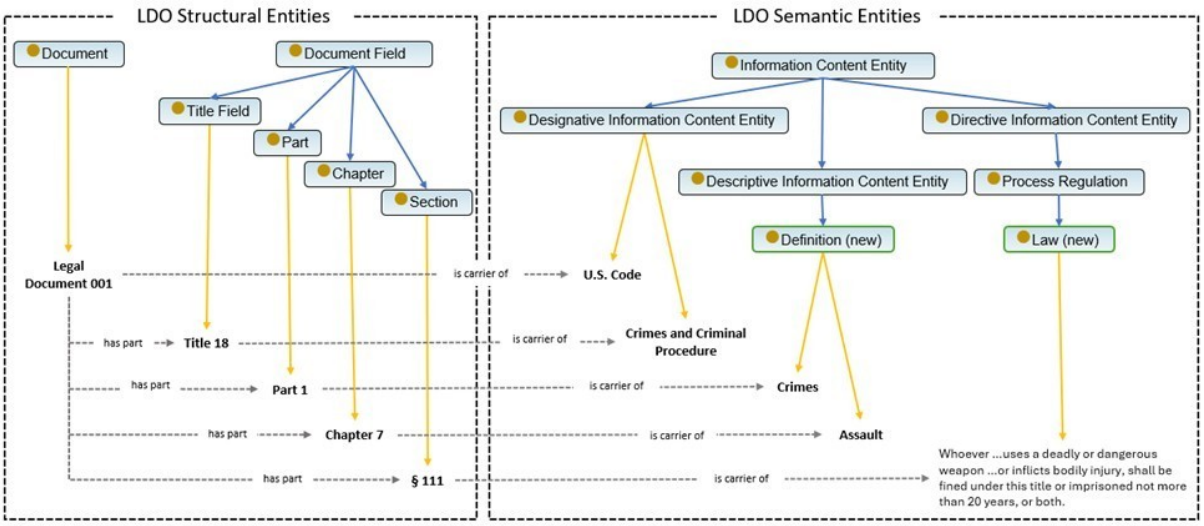

Figure 3. A visual analysis of the structural and semantic units in a Title, Part, Chapter, and Section of the U.S. Code. The Title carries a designation of Crimes and Criminal Procedure, the Part and Chapter carry a Definition for Crimes and Assault, respectively, and the diagrammed Section carries a Law stating "whoever...uses a deadly or dangerous weapon… or inflicts bodily injury, shall be fined under this title or imprisoned not more than 20 years, or both."

In effect, the structural units of a legal clause or subsection are semantically loaded, cross-referenced, and hierarchically interrelated. This makes the U.S. Code ideal for formal modeling in legal informatics, semantic web technologies, and regulatory AI systems. Each Legal Document in the U.S. Code is an Information Bearing Artifact (IBA) that is designed to bear some specific Information Content Entity (ICE). Each ICE, in turn, generically depends on a series of paragraphs of text or diagrams in the form of physical pieces of paper or an electronic word processor file.

The LDO could help law enforcement operations by transforming thousands of pages of legal text into an integrated, machine-readable, logically connected knowledge base that officers, analysts, and automated systems can query in real time. Such a knowledge base would facilitate new capabilities that include:

1. Rapid Legal Reference in the Field: Law Enforcement Officers often rely on printed materials or generic keyword searches to find relevant laws or regulations. In contrast, an LDO compliant knowledge base would provide linked citations between statutes, regulations, and agency policies.

2. Cross-Jurisdictional Clarity: An LDO compliant knowledge base could resolve overlapping federal, state, and local authorities, thereby providing clarity for Law Enforcement Officers.

3. Automated Case Classification: Reports and evidence must be matched to the correct legal codes for prosecution. An LDO complaint knowledge base would provide automated suggestions for charges, penalties, and procedural requirements with case data.

4. Threat & Pattern Detection: Emerging threats often cut across multiple legal domains. An LDO complaint knowledge base would support legal knowledge graphs that reveal multi-domain violations and coordinated criminal activity.

5. Training & Knowledge Transfer: New personnel face steep learning curves learning legal frameworks. An LDO complaint knowledge base would reduce training time by making legal relationships visual and intuitive.

6. Decision Support & Compliance: Law Enforcement Officers need to know *exactly* what they can and cannot do under law. Encoding rules of engagement and procedural constraints directly linked to statutory authority would allow for immediate decision support — e.g., "under USC Title X, section Y, you may detain for… but require a warrant for…"

7. Integration with Sensors & Intelligence Systems. Sensors may detect anomalies but lack legal context. An LDO complaint knowledge base could be used to link event data (e.g., vessel location, drone flight path) to applicable legal zones and restrictions.

In summary, the LDO will make any body of laws computable and navigable, turning it into an operational asset rather than a static reference. This empowers law enforcement to act faster, more accurately, and in legal alignment — while also supporting intelligence fusion, prosecution, and prevention.

6. Conclusion

To represent the evolving nature of law, an ontology must do more than capture legal terminology —it must capture legal meaning. The Legal Document Ontology (LDO) introduced here enables such expressivity by formally distinguishing between structural and semantic entities within legal texts. By aligning designative, descriptive, and directive content with structural components such as clauses and sections, the LDO allows for a nuanced, machine-interpretable account of legal reasoning and authority.

This dual-axis approach—structural and semantic—facilitates not only machine interpretation but also human comprehension by clarifying how legal meaning emerges from document architecture. The integration of the Common Core Ontologies with insights from the Document Components Ontology and Information Artifact Ontology provides a framework capable of modeling legal content far beyond static textual references.

The Common Core Information Entity Ontology offers the clearest mapping of information content, but it needs an equally precise account of information structure. We advocate such a development by joining the plethora of relevant classes in the Document Component Ontology and the Information Artifact Ontology with the Common Core hierarchy. This would create a maximally representative and machine-interpretable picture of law in contexts besides the U.S. Code. In a digital and increasingly automated legal landscape, such formal representation is essential.

Through the LDO, we propose a principled method for parsing, integrating, and reasoning over legal corpora—one that is faithful to both the structure and the action-oriented nature of law. This enables legal systems to scale in complexity while remaining accessible and intelligible. Modeling legal knowledge must go past mere reference to titles or paragraphs of text. Instead, an effective alternative might consider the actions that a law prohibits, permits, or requires.

Acknowledgements

This work was completed under the auspices of CUBRC, Inc. Cameron More, Carter Beau-Benson, Adam Czerniejewski, Mark Jensen, and the members of the National Center for Ontological Research provided valuable assistance and critique at every stage of development.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] D. Jurafsky and J. H. Martin. Speech and Language Processing (3rd ed., draft). Stanford University, 2023. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/.

[2] T. R. Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2):199–220, 1993. [Online]. Available: https://doi.org/10.1006/knac.1993.1008.

[3] R. Feldman and J. Sanger. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2007.

[4] C. Bizer, T. Heath, and T. Berners-Lee. Linked data—The story so far. International Journal on Semantic Web and Information Systems, 5(3):1–22, 2009. [Online]. Available: https://doi.org/10.4018/jswis.2009081901

[5] R. Arp, B. Smith, A. D. Spear, Building Ontologies with Basic Formal Ontology, MIT Press, Cambridge, Massachusetts, 2015.

[6] J. N. Otte et. al., BFO: Basic Formal Ontology, Applied Ontology, 17(1), 17–43, 2022.

[7] Common Core Ontologies, 2025. URL: https://commoncoreontology.github.io/cco-webpage/.

[8] Jensen, M., De Colle, G., Kindya, S., More, C., Cox, A. P., & Beverley, J. (2024). *The Common Core Ontologies*. In Formal Ontology in Information Systems (Vol. 394, pp. 47–58). Frontiers in Artificial Intelligence and Applications.

[9] J. L. Austin, *How to Do Things with Words*. J. O. Urmson and Marina Sbisà (Eds.). Oxford University Press, Oxford, England, 1962.

[10] *United States v. Carolene Products Co.*, 304 U.S. 144, 1938. [Online]. Available: https://supreme.justia.com/cases/federal/us/304/144/

[11] Information Artifact Ontology (IAO), 2025. URL: http://purl.obolibrary.org/obo/iao.owl.

[12] A. Constantin, S. Peroni, S. Pettifer, D. Shotton, and F. Vitali. The Document Components Ontology (DoCO). Semantic Web, 7(2):167–181, 2016. [Online]. Available: http://dx.doi.org/10.3233/SW-150177

[13] Hume, D. (1978). *A treatise of human nature* (L. A. Selby-Bigge & P. H. Nidditch, Eds.). Oxford University Press. (Original work published 1739–1740)