VARONA: Verifiable Analysis of Regulatory Objects, Narratives, and Actors

AI-powered information retrieval and knowledge networks for regulatory readiness

Draft v1.0 | 2026-08-18 | Working Paper

Edwin Alvarado-Mena, PhD

Policy Data Science Consulting LLC

AlvaradoCSS.com ↗ | PolicyDataScience.com ↗

ABSTRACT

Policy documents contain structured assertions about problems, interventions, institutional responsibilities, affected actors, expected consequences, and evidence. Yet conventional text representations based on topics, named entities, or co-occurrence often discard these relationships. This manuscript introduces VARONA (Verifiable Analysis of Regulatory Objects, Narratives, and Actors), an ontology-guided system for transforming unstructured policy documents into proposition-centered knowledge networks. VARONA treats policy propositions as first-class analytical objects and connects normalized entities to them through explicit semantic roles, including source, authority, responsible actor, instrument, affected actor, beneficiary, cost bearer, risk bearer, and location. Proposition-level attributes preserve predicates, policy status, modality, expected and adverse effects, evidence basis, and extraction confidence without unnecessarily expanding graph topology. The system combines large-language-model extraction with deterministic processing, document provenance, and verbatim evidence so that extracted relationships remain traceable to their textual sources. A two-pass procedure separates proposition identification from proposition structuring, reducing the risk that ontology vocabulary induces unsupported propositions. The resulting representation supports multiple analytical projections of the same network, including institutional, distributional, stakeholder, evidentiary, and substantive views. An ongoing application to Arizona data-center policy demonstrates the architecture and its use as both an analytical interface and a human-in-the-loop environment for identifying extraction errors and refining the protocol. VARONA provides a framework for moving from policy-text retrieval to structured, inspectable policy analysis while preserving the documentary evidence underlying the resulting knowledge representation.

Keywords: policy data science; regulatory readiness; knowledge graphs; large language models; information retrieval; ontology engineering; data centers

Contents

One-Pager

Policy documents encode substantive accounts of how problems are understood and addressed, who is involved and affected, and what consequences are expected to follow. Yet this information is largely embedded in unstructured prose, making systematic comparison difficult and limiting actors' ability to navigate document overload in complex policy environments.

Common text representations based on topics, sentiments, named entities, or co-occurrence can identify what documents discuss and which actors appear together. They do not, however, reliably preserve the structure of policy propositions. This limitation is particularly consequential in contested and rapidly evolving policy domains. In the data center industry, for example, knowing that a regulator, data-center operator, and ratepayer group appear in the same passage does not reveal the policy interplay among them.

VARONA addresses this problem by transforming policy documents into proposition-centered knowledge networks. A policy proposition is a discrete assertion that expresses a policy-relevant condition, action, relationship, position, or consequence and identifies how relevant entities participate in that assertion. VARONA's ontology-guided extraction pipeline identifies user-defined policy propositions, normalizes relevant entities, assigns them to explicit roles within each proposition, and preserves the source text supporting every extraction.

Consider a recommendation that a regulator require large-load customers to pay for the grid upgrades needed to serve them, thereby protecting existing ratepayers from cost shifting. A co-occurrence network might simply connect the regulator, large-load customers, and ratepayers because they appear in the same passage.

VARONA instead represents the recommendation itself as a proposition node. The regulator can be linked to the proposition as an authority or responsible actor, a tariff as the policy instrument, large-load customers as cost bearers, and existing ratepayers as beneficiaries. The network therefore preserves how each entity participates in the policy claim rather than reducing their relationships to generic associations.

Information that describes the proposition itself—such as its source-faithful predicate, normalized action, policy status, expected effects, evidentiary basis, and extraction confidence—is stored as proposition-level information rather than converted into additional graph nodes. This preserves analytical detail without making the network unnecessarily complex.

Because these relationships are encoded within the same underlying network, users can examine the policy landscape from different perspectives. An institutional view can reveal authority and implementation responsibility; a distributional view can reveal who is expected to benefit, bear costs, or face risks. Other views can focus on stakeholder positions, evidence, or particular policy domains.

Throughout these analyses, propositions remain linked to verbatim evidence and document provenance, allowing users to inspect the source material behind the structured representation rather than treating model-generated structure as self-validating data.

VARONA currently produces structured proposition data, an interactive network for policy analysis, and a data model that can support reproducible automated reports. Its current implementation applies the approach to Arizona data-center policy and demonstrates how proposition-centered extraction can turn complex regulatory documents into a structured, inspectable policy knowledge system.

VARONA is part of The Manifold — AI for Policy Data Science and can be adapted to other policy domains.

1 Introduction

Policy analysis depends heavily on text. Statutes, regulations, reports, public comments, and countless other sources describe how problems are understood and how public- and private-sector actors propose to address them (Grimmer, Roberts, and Stewart 2022; Sewerin et al. 2023). Policy-relevant documents contain information about the causes and consequences of policy problems, the institutions and instruments involved in addressing them, the distribution of benefits and burdens, and the evidence and disagreements that shape policy choices.

Actors need to make the most of the information available in source documents to improve strategy, planning, and execution. The difficulty is not simply finding relevant documents or processing large collections. The actual challenge is to recover as much useful structure as possible from the information these sources contain. A single information-rich document, or a small set of documents, can require the same kind of systematic extraction and analysis as a much larger corpus.

Consider a regulatory document and, more specifically, a statement recommending that a regulator require large electricity customers to assume infrastructure costs associated with their extraordinary demand. I will call this a policy proposition. Policy propositions can be represented in several ways. A keyword representation might identify terms such as *regulator*, *electricity*, *customers*, *infrastructure*, and *costs*, facilitating filtering. Named-entity recognition might identify the regulator and utility. A co-occurrence network might connect those named entities because they appear in the same paragraph. A topic model might classify the passage under electricity regulation, infrastructure finance, or any other category (Eisenstein 2019; Grimmer, Roberts, and Stewart 2022).

All of these representations can be useful across multiple applied settings, but none necessarily preserves the policy proposition itself. The analytically important structure is closer to:

- An actor has responsibility or authority.
- A particular action is recommended or described.
- The action may operate through a policy instrument.
- The intervention has a target.
- Different actors may be affected.
- Some actors may benefit.
- Others may bear costs or risks.
- The proposition may be an existing policy, a recommendation, an implementation action, a causal claim, or a stakeholder position.
- The proposition may have an explicit evidentiary basis.
- Every extracted claim should remain traceable to the language from which it was derived.

This emphasis on recovering the full structure of policy propositions motivates VARONA. A policy proposition is a discrete assertion that expresses a policy-relevant condition, action, relationship, position, or consequence and identifies how relevant entities participate in that assertion. VARONA uses this unit to structure information that would otherwise remain embedded in regulatory text.

VARONA is an ontology-guided system for constructing and inspecting **policy knowledge networks** from unstructured documents using large language models (LLMs) (Ziems et al. 2024; Hogan et al. 2021; Kommineni, König-Ries, and Samuel 2024). Its central representational choice is to treat propositions, rather than entity co-occurrence, as the semantic center of the graph. Named entities connect to propositions through explicitly typed roles. Other useful qualifications remain properties of the proposition and are available through an interactive inspector. Once the LLM retrieves relational data grounded in the ontologies of interest, the resulting network can therefore preserve distinctions that disappear in conventional actor networks.

Importantly, VARONA is not intended to make every element of an ontology visible simultaneously. Such a design would quickly produce an unreadable graph. Instead, VARONA deliberately separates three functions:

1. **Graph topology** represents propositions, entities, and analytically meaningful semantic roles.
2. **Proposition attributes** preserve qualifications that do not need to become independent nodes.
3. **Interactive projections** selectively expose different dimensions of the same underlying representation.

While actors need to make the most of the information available in their source documents, they also must ensure that the LLM-retrieved policy propositions are reliable. Therefore, VARONA retains verbatim evidence and document provenance. These safeguards are important for substantive interpretation, but they serve a second purpose: VARONA functions as an inspection environment for the information-extraction process itself. Implausible thresholds, malformed targets, missing participant roles, overly broad problem statements, and other extraction problems become visible when structured information is placed beside its textual evidence.

This draft manuscript describes VARONA's conceptual and computational architecture. It showcases an ongoing application to Arizona data-center policy as a development case. Because the extraction protocol continues to be refined and some imprecisions remain, the application is presented as a demonstration of representation and analytical capability rather than as a finalized empirical study of Arizona policy.

VARONA contributes a proposition-centered alternative to entity co-occurrence representations of policy documents. It operationalizes a policy ontology in which claims, actions, participant roles, evidence, and domain-specific attributes remain distinguishable within one underlying network. Multiple projections make this structure analytically accessible, while provenance and extraction diagnostics allow users to verify the model-generated representation against its source.

2 From Policy Documents to Policy Knowledge

2.1 The limits of document-level representation

A document is a container of natural language, which is not necessarily the appropriate unit of analysis for examining regulatory landscapes. Imagine a taskforce that produces a report to a senior decision-maker. A single report can contain hundreds of distinct propositions: descriptions of existing conditions, forecasts, recommendations, implementation actions, statements of authority, expressions of support or opposition, quantitative estimates, and so on.

Now imagine a collection of policy-relevant documents, each containing hundreds of distinct policy propositions. Given a set of documents, a document-level classification protocol can answer questions such as:

- Which documents discuss water?
- Which reports mention transmission?
- Which jurisdictions discuss data centers?
- Which organizations appear frequently?

These questions are critical across multiple settings. Document-level analyses are certainly valuable. That said, many regulatory readiness questions require a more granular representation to answer strategic questions such as:

- Who is expected to implement a recommendation?
- Which actor is supposed to pay?
- Who is expected to benefit?
- Which propositions invoke regulatory authority?
- What policy instruments are associated with water conservation?
- Which stakeholder positions oppose a particular intervention?
- Which claims are forecasts rather than observed facts?
- What evidence supports a proposed policy?
- Where do actors disagree about the consequences of the same intervention?

Answering these questions requires recovering relations (structure) inside policy propositions, not mere document-level classification.

2.2 Entity co-occurrence is not policy-proposition structure

Actor and entity networks often infer relationships from shared documents, paragraphs, or sentences. Such networks can reveal patterns of attention and association, making them key tools for regulatory readiness (Zufall and Scott 2024).

But co-occurrence is semantically ambiguous. If a regulator, utility, data-center operator, and ratepayer organization occur in the same paragraph, their co-occurrence does not tell us whether the regulator regulates the utility, whether the utility imposes costs on the data center, whether the data center benefits ratepayers, or whether the ratepayer organization opposes a tariff.

VARONA therefore does not treat co-occurrence as evidence of a specific policy relationship. Rather, VARONA preserves the proposition-level semantics that establish the relationship and reveal its policy-relevant structure.

2.3 Why policy propositions should be first-class objects

A policy proposition is useful as an intermediate representational unit because multiple entities can occupy different roles within the same assertion. Following our initial example, a simplified policy proposition may contain:

```
Proposition type: RECOMMENDATION
Predicate: require large-load customers to pay for grid upgrades
Controlled action: REQUIRE
Responsible actor: regulator
Instrument: tariff
Target: large-load customer
Beneficiary: existing ratepayers
Cost bearer: large-load customer
Expected effect: reduce cost shifting
Status: recommended
Evidence: verbatim source passage
```

Flattening this proposition into a single entity-to-entity edge would discard most of the critical information that could be recovered from text. VARONA approaches this task by representing the proposition itself as a node.

Creating separate direct edges for every possible relationship can also become problematic because qualifications such as modality, status, evidence, and provenance may apply to the proposition as a whole rather than to one pairwise edge. VARONA, however, embeds this information as node properties.

3 Design Principles

VARONA is thus a system organized around seven design principles.

3.1 Ontology-guided extraction

The system does not ask a language model to produce an unconstrained, open-ended information summary or an arbitrary knowledge graph. Quite the opposite, LLM-powered extraction is guided by predefined entity classes, proposition types, controlled actions, semantic roles, status categories, evidence categories, and domain attributes. The proposition's source-faithful predicate is preserved separately from this controlled vocabulary.

All these elements pertain to the ontology of interest. Grounded in policy expertise, ontologies serve as a contract between extraction and analysis (Keet 2018; Tudorache 2020).

3.2 Proposition-centered representation

The system assumes that policy propositions are first-class analytical objects. They are represented as nodes in the policy knowledge network. Entities participate in propositions through typed semantic roles rather than being connected merely because they occur together.

3.3 Evidence grounding

LLM-extracted structure should remain auditable against source language (Törnberg 2024; Wang et al. 2024). VARONA retains verbatim evidence and document location so users can move from graph abstraction back to the input text. The interface also shows the previous and next paragraphs.

3.4 Separation of topology and attributes

Not every ontology element should become a node or edge. Participant ontology roles such as `BENEFICIARY`, `AUTHORITY`, or `COST_BEARER` are naturally represented as graph structure because they specify distinct ways in which entities participate in a policy proposition. VARONA distinguishes four complementary ways of referring to the same relation: a natural-language label used in prose, a canonical-schema label used for stable machine-facing representation, an ontology-role label naming the entity's semantic function within a proposition, and a verb-like relational predicate used in the interactive network. For example, entities described in prose as **beneficiaries** are stored under the canonical-schema label `BENEFICIARIES`, occupy the ontology role `BENEFICIARY`, and are rendered in the network through the relational predicate *benefits from*. This separation preserves stable computational representations and semantic role categories while keeping the graph readable as relational statements.

Properties such as predicate, controlled action, confidence, modality, status, evidence basis, and many domain-specific attributes are generally better represented as proposition properties. This allocation avoids an unnecessarily dense graph.

3.5 Multiple projections, one underlying representation

Different policy questions require different views of the same networked data. VARONA supports analytical lenses rather than constructing unrelated networks for every question. A menu of these network views is implemented directly in the interface.

3.6 Human inspectability

LLM output is not treated as ground truth merely because it has been converted into structured data (Ashwin, Chhabra, and Rao 2023; Bisbee and Spirling 2026). The system is designed to make extraction decisions visible and questionable values inspectable. The interface also preserves a context window around each policy proposition, allowing users to inspect where and how it appears in the source text, and the respective verbatim excerpts and rationales for extraction.

3.7 Conservative treatment of missing information

VARONA avoids interpreting missing fields as substantive negatives unless the extraction protocol explicitly supports that interpretation. It follows that an absent value does not necessarily mean that a concept is absent from the real-world policy situation. Missingness may mean that the concept was not applicable, was not stated, or was not successfully extracted.

4 Ontological Representation

Rather than relying on a single flat taxonomy, VARONA integrates several interlocking ontological layers.

Although the current implementation operationalizes these layers for the data center industry, VARONA is domain-adaptable and can operate with ontologies tailored to other policy contexts.

4.1 Policy-domain ontology

The first ontological layer identifies the substantive policy issue addressed by a proposition. The domain ontology answers:

> What substantive policy issue is this proposition about?

In the current data-center implementation, VARONA distinguishes broad domains concerning data-center demand, energy and grid systems, water and cooling, siting and permitting, enabling infrastructure, economics and workforce, and community and environmental impacts. These areas are represented through substantive proposition types and domain-specific attributes rather than through a single flat `domain` field.

For example, a `GRID_CLAIM` identifies a proposition about energy or grid systems, while energy-specific attributes can preserve additional information about the particular issue addressed. The same logic applies to `WATER_CLAIM`, `SITING_CLAIM`, `INFRASTRUCTURE_CLAIM`, `ECONOMIC_CLAIM`, `DEMAND_CLAIM`, and `COMMUNITY_IMPACT`.

This system design allows VARONA to identify the broad substantive area of a proposition while retaining more specific information within that area. Substantive domains should be distinguished from other dimensions of the ontology that can apply across them. A recommendation, for example, can concern water conservation, electricity rates, permitting, or workforce development. Likewise, questions of institutional authority, implementation, stakeholder position, evidence, and distributional consequences can arise within any substantive domain.

VARONA therefore does not treat these dimensions as additional members of a single domain taxonomy.

4.2 Proposition types

A second ontological layer identifies what kind of policy assertion is being made. The proposition type captures what the statement is doing, while the substantive domain, as seen above, captures what it is about.

The current proposition vocabulary distinguishes substantive claims from propositions that perform other policy functions, including recommendations, descriptions of existing policy, implementation actions, authority claims, evidence claims, stakeholder positions, and tradeoffs.

This distinction allows substantive content and propositional function to be represented separately. A proposition can therefore concern water and cooling while functioning as a recommendation, or concern energy and grid systems while describing an existing policy or implementation action.

4.3 Cross-cutting policy dimensions

Other information cuts across both substantive domains and proposition types:

- Institutional roles identify who has authority, responsibility, or another defined role in a proposition.
- Distributional roles distinguish who is affected, who benefits, who bears costs, and who bears risk.
- Stakeholder-position information captures explicitly attributed support, opposition, concern, dissent, or alternatives.
- Evidence and uncertainty characterize the basis and epistemic status of policy claims.
- Implementation-related information can indicate policy status, conditions, timing, or thresholds.

These dimensions are not separate policy domains. They provide additional structure for understanding propositions within and across substantive areas.

This distinction also underlies VARONA's interactive analytical projections (previously called lenses): institutional, distributional, political, and evidentiary views reorganize the same underlying propositions according to different cross-cutting dimensions, while substantive views focus on areas such as energy, water, siting, or economics.

4.4 Entity ontology

The entity layer defines the kinds of actors and policy objects that can participate in propositions. Current classes include organizational and institutional actors relevant to the data center industry, such as:

- Data-center operators.
- Hyperscalers.
- Colocation providers.
- Cloud or AI companies.
- Developers.
- Utilities.
- Grid operators.
- Government actors.
- Regulatory actors.
- Tribal governments.
- Local governments.
- Community actors.
- Civil-society actors.
- Knowledge actors.
- Investors or financiers.
- Suppliers or contractors.

The ontology of entities also represents non-organizational policy objects, including:

- Legal authorities.
- Policy instruments.
- Programs or projects.
- Energy technologies.
- Cooling or water technologies.
- Infrastructure assets.
- Policy outputs.
- Places.
- Formal roles.
- Policy bodies.
- Stakeholder categories.
- Affected populations.

Entity normalization assigns canonical identities so repeated surface mentions can be connected to the same object across propositions. This is critical to maintaining a system of meaningful, cumulative information.

4.5 Proposition ontology

The proposition layer specifies the kinds of policy assertions VARONA can represent.

The current vocabulary includes:

- `DEMAND_CLAIM`.
- `GRID_CLAIM`.
- `WATER_CLAIM`.
- `SITING_CLAIM`.
- `INFRASTRUCTURE_CLAIM`.
- `ECONOMIC_CLAIM`.
- `COMMUNITY_IMPACT`.
- `RECOMMENDATION`.
- `EXISTING_POLICY`.

- `IMPLEMENTATION_ACTION`.
- `AUTHORITY_CLAIM`.
- `EVIDENCE_CLAIM`.
- `STAKEHOLDER_POSITION`.
- `TRADEOFF`.

This vocabulary deliberately combines substantive claims with policy-process propositions while preserving analytically useful distinctions among them. A forecast about electricity demand is analytically different from a recommendation to modify a tariff, a statement about regulatory authority, or an organization's opposition to a proposed rule.

4.6 Predicate and controlled action vocabulary

Every structured proposition preserves a compact, source-faithful `predicate` describing what the proposition actually asserts. The predicate is not constrained to a closed vocabulary and is intended to retain meaning that would be distorted by forcing every proposition into a predefined action class.

VARONA also provides an optional controlled `action` as a secondary normalization safeguard. It is populated only when the proposition's operative predicate can be represented faithfully by the controlled vocabulary; otherwise it remains null. This distinction is especially important for descriptive or evidentiary propositions such as *is changing rapidly* or *accounts for 27% of generation*, which have meaningful predicates but need not correspond to a policy operation.

The current controlled action vocabulary includes:

- `IDENTIFY`.
- `DEFINE`.
- `FORECAST`.
- `CAUSE`.
- `CONTRIBUTE_TO`.
- `INCREASE`.
- `REDUCE`.
- `EXACERBATE`.
- `CONSTRAIN`.
- `DELAY`.
- `CREATE_RISK`.
- `RECOMMEND`.
- `ESTABLISH`.
- `REQUIRE`.
- `AUTHORIZE`.
- `PROHIBIT`.
- `APPROVE`.
- `REGULATE`.
- `ZONE`.
- `PERMIT`.
- `FUND`.
- `PAY`.
- `INCENTIVIZE`.
- `TAX`.
- `REPORT`.
- `DISCLOSE`.
- `MONITOR`.
- `EVALUATE`.
- `PLAN`.
- `COORDINATE`.

- `CONSULT`.
- `NEGOTIATE`.
- `PROCURE`.
- `CONNECT`.
- `BUILD`.
- `DEVELOP`.
- `UPGRADE`.
- `DEPLOY`.
- `ACCELERATE`.
- `STREAMLINE`.
- `IMPROVE`.
- `EXPAND`.
- `ENCOURAGE`.
- `FACILITATE`.
- `ENABLE`.
- `CURTAIL`.
- `CONSERVE`.
- `REUSE`.
- `PROTECT`.
- `BENEFIT`.
- `BURDEN`.
- `IMPOSE_COST`.
- `SHIFT_COST`.
- `ASSIGN_RESPONSIBILITY`.
- `SUPPORT`.
- `CONDITIONALLY_SUPPORT`.
- `EXPRESS_CONCERN`.
- `DISSENT`.
- `OPPOSE`.
- `PROPOSE_ALTERNATIVE`.
- `BALANCE`.

The controlled vocabulary makes comparable policy operations easier to aggregate across differently worded passages without requiring every proposition to be reduced to one of these actions.

4.7 Proposition-mediated policy relations (relationship roles)

The underlying logic is that entities participate in propositions through typed semantic roles that correspond to substantively different policy relations.

VARONA distinguishes ontology roles from their canonical-schema representation and from the labels used in prose and visualization. Ontology roles are singular noun-like categories that identify the semantic function occupied by an entity within a proposition. Canonical-schema labels provide the machine-facing representation of those roles and may be plural when a field can contain multiple entities. Natural-language labels keep the manuscript readable, while the interactive network expresses the same relations as verb-like predicates. The distinction is representational rather than substantive: an entity described in prose as a **beneficiary** is stored under `BENEFICIARIES`, occupies the ontology role `BENEFICIARY`, and appears on an edge labeled *benefits from*.

The ontology roles are:

- `SOURCE` — attribution: the actor-like entity that is explicitly presented as the communicative or epistemic originator of the proposition.
- `INSTRUMENT` — policy instrumentation: the legal, regulatory, programmatic, or administrative mechanism through which the proposition operates.

- `AUTHORITY` — institutional authority: the actor or legal authority that explicitly possesses jurisdiction, approval power, or governing authority relevant to the proposition.
- `RESPONSIBLE_ACTOR` — implementation responsibility: the actor explicitly assigned to carry out, administer, decide, regulate, report, monitor, build, coordinate, or otherwise implement the proposition.
- `AFFECTED_ACTOR` — policy exposure or impact: an actor or population explicitly subject to, regulated by, served by, or otherwise affected by the proposition.
- `BENEFICIARY` — distribution of benefits: an actor or population explicitly expected to receive a benefit.
- `COST_BEARER` — distribution of costs: an actor or population explicitly expected to pay, finance, absorb, or shoulder a cost.
- `RISK_BEARER` — distribution of risk: an actor or population explicitly exposed to a stated risk or adverse uncertainty.
- `LOCATION` — spatial applicability or context: a place or jurisdiction explicitly associated with the proposition.

These are more than generic graph edges. They encode institutional, instrumental, attributional, distributional, impact, and spatial phenomena asserted in the underlying text. Formally, VARONA implements them as entity-to-proposition relations because the proposition is the unit that binds their meanings together.

4.8 Proposition qualifications

Additional fields characterize the proposition without necessarily becoming graph relationships. These include:

- Problem.
- Objective.
- Target.
- Data-center segment.
- Policy status.
- Modality.
- Condition.
- Time.
- Threshold.
- Expected effects.
- Adverse effects.
- Evidence basis.
- Stakeholder position.
- Extraction confidence.
- Domain-specific energy, water, siting, and economic attributes.

The epistemic vocabulary can distinguish evidence such as empirical facts, quantitative estimates, forecasts, scenarios, comparative cases, legal text, stakeholder judgment, and uncertainty.

5 Pipeline Modules

VARONA operates through an ontology-guided information-retrieval workflow. The workflow is a modular pipeline.

The current system design separates document processing into stages so intermediate representations can be inspected and revised independently.

5.1 Stage 01: document parsing

Source material is converted into paragraph-level records with document and location identifiers, facilitating programmatic manipulation. Paragraph-level source traceability provides the basic unit for tracing later extractions back to the source text.

5.2 Stage 02: policy-information detection

Paragraphs are evaluated through a recall-oriented screen for potentially relevant policy information. This stage helps locate where broad policy domains may be represented, but it is not used as a proposition-existence gate.

Because the screen favors recall, later stages must still distinguish substantive assertions from lexical or contextual matches.

5.3 Stage 03: entity extraction

Policy-relevant entities are identified. The extraction protocol includes named organizations and institutions as well as bounded generic actors and affected populations when these are analytically meaningful.

5.4 Stage 04: entity normalization

Entity mentions are resolved into a canonical registry. Normalization is essential for network construction because textual, surface variants should not automatically become separate graph nodes.

5.5 Stage 05: policy-information recovery

The project also contains a policy-information recovery stage for structured evidence coding. This is an analytical product of the broader pipeline rather than a required serial input to Stage 06 in the current architecture.

5.6 Stage 06: proposition extraction

Stage 06 converts policy text into structured proposition frames while using the normalized entity registry to populate semantic participant roles (Liu et al. 2020; Xie et al. 2022; Brach, Košťál, and Ries 2025). The current implementation separates this stage internally into proposition identification and proposition structuring.

5.6.1 Proposition identification

The first pass asks only whether a source record contains one or more genuine policy propositions and, when it does, recovers the minimum verbatim evidence span for each distinct proposition. It does not assign proposition types, controlled actions, participant roles, domain attributes, evidence categories, or extraction confidence.

Before this pass, VARONA now constructs a conservative analysis-text layer from the Stage-01 records. This layer does not rewrite the parsed source. Instead, when a body or list-item record ends syntactically incomplete and the next substantive text clearly continues it, the preparation code joins the split text for proposition analysis.

VARONA can bridge intervening bibliographic records and retain a narrow substantive tail when a PDF parser has placed the beginning of the next sentence after a citation on the same extracted line. The first source paragraph remains the anchor identifier, while the contributing paragraph and page identifiers are retained in the prepared record. This system design addresses page-boundary artifacts without erasing original provenance or requiring entity normalization to be rebuilt solely because one source paragraph was physically split by PDF layout.

Separating proposition identification from the larger ontology reduces the risk that policy vocabulary itself creates false propositions. Bibliographic references and a narrow set of unambiguous document-status artifacts can also be removed deterministically before this pass.

Formatting alone is not an exclusion criterion: a heading, numbered recommendation, or list item is retained when it contains a complete substantive assertion. Proposition identification also requires a grammatically complete span; abbreviations, citation markers, and page-fragment prefixes or suffixes are not treated as proposition boundaries.

The output of this pass is therefore an explicit intermediate representation of the proposition itself. When a paragraph contains several assertions, they are recovered as separate proposition-local evidence spans rather than one paragraph-wide frame.

5.6.2 Proposition structuring

The second pass structures one already-identified proposition at a time. It receives the exact proposition span, the reconstructed analysis context when needed, the contributing source paragraph/page identifiers, and only the normalized entities mentioned in those contributing source records rather than the document-wide registry.

This changes the role of context. The proposition span is authoritative for proposition meaning and participant roles, and surrounding text (paragraphs) is used only when necessary to resolve an explicit antecedent, attribution, or scope. A broader or narrower actor category mentioned elsewhere in the analysis context therefore does not automatically inherit a role in the focal proposition.

The structuring pass assigns the proposition type, a compact source-faithful predicate, participant roles, modality, consequences, evidence basis, domain-specific attributes, and other proposition qualifications. Controlled `action` is a secondary normalization and is optional: it is populated only when the controlled vocabulary faithfully represents the proposition's operative predicate.

Policy or implementation `status` is likewise nullable for descriptive claims. In an explicit recommendation or imperative, the head verb of the recommended act receives priority when a controlled action is appropriate; an intended consequence, nearby policy noun, or modal embedded in a subordinate clause does not replace the operative predicate. The structuring pass also performs explicit proposition-local checks for attribution, execution responsibility, institutional authority, and directional benefit/cost/risk roles.

Stage 06A owns proposition admission. Once a proposition packet reaches 06B, the current schema requires a structured frame rather than allowing 06B to apply a second broad null/rejection gate.

A deterministic finalization step then restores authoritative provenance and evidence from the prepared packet, utilizes canonical entity names/types from Stage 04, removes type-incompatible role assignments without inferring replacements, and restores the canonical Stage-06 output schema used by VARONA.

Paragraphs with no Stage-06A propositions remain present with empty `frames` lists, preserving downstream compatibility and source traceability.

This two-pass design keeps Stage 02 as a recall-oriented screen and Stage 05 as a parallel policy-information recovery product rather than making either a serial prerequisite for proposition extraction.

The critical dependency for Stage 06 is instead the proposition-local evidence span together with the Stage-04 normalized entity registry.

5.6.3 Reproducible API execution and batching

The current implementation automates the two model-dependent passes through a scripted Python runner rather than requiring manual transfer of prompt inputs and outputs.

Since VARONA uses proprietary models, it includes safeguards against unintended or unwise API calls. The Python runner partitions analysis records and proposition packets into bounded batches, executes model calls through Structured Outputs, validates returned records against prepared inputs, and persists both whole-batch and item-level caches so interrupted or partially complete runs can resume without paying again for successful items.

Before any new paid requests are issued, a cost-control preflight reports reusable cached items, new items, planned paid requests, and a conservative estimate of additional spend. It also requires explicit user confirmation unless intentionally bypassed.

To keep the system cost-conscious, deterministic code performs all operations that do not require semantic interpretation. These include citation and document-status exclusion, conservative cross-page reconstruction, proposition identifiers, paragraph-local entity selection, longest-match resolution of overlapping entity surfaces, authoritative provenance/evidence restoration, canonical entity-name/type injection from Stage-04 entity IDs, entity-role type compatibility checks, and final assembly of the downstream Stage-06 artifact.

In the API representation, the structuring model selects participant entity identifiers rather than regenerating canonical entity metadata; the latter is restored programmatically from the normalized registry. Model calls are therefore reserved for proposition identification and semantic proposition structuring.

The current development configuration uses `gpt-5.6` with low reasoning for 06A, preserving already-paid reusable proposition-identification results, and `gpt-5.4-nano` with low reasoning for high-volume 06B structuring. `gpt-5.6-luna` is configured as an optional escalation model, but low-confidence escalation is disabled by default. These choices are execution settings rather than elements of the ontology and can be changed independently of the canonical Stage-06 schema.

The batching layer embodies an execution strategy: each 06A record and each 06B proposition packet retains the same analytical meaning it would have in an individual request. In the current configuration, synchronous requests are bounded by both item count and serialized character count (25 records/120,000 characters for 06A and 50 packets/350,000 characters for 06B). The same runner can use synchronous resumable requests or the asynchronous Batch API without changing the canonical Stage-06 schema.

This division of responsibility is important: 06A determines whether a proposition exists, 06B structures admitted propositions, and deterministic code protects identities, provenance, evidence, and mechanically enforceable ontology constraints.

6 Graph Model

6.1 Two primary node classes

VARONA's graph contains two central node classes (Hogan et al. 2021; Barrasa and Webber 2023):

1. **Entity nodes** represent canonical actors and policy objects.
2. **Proposition nodes** represent extracted policy propositions.

Proposition nodes can be visually distinguished from entity nodes so the basic grammar of the network remains legible.

6.2 Typed proposition-mediated policy relations

Edges connect entities to propositions through typed policy relations. These relations are source-grounded insofar as the corresponding actor or policy object, its role, and the proposition are explicitly supported by the source passage and correctly extracted.

As mentioned earlier, VARONA represents actual relational phenomena asserted in policy text rather than inferring a relationship merely from co-occurrence.

The same relationship can therefore be named differently depending on whether the manuscript is referring to prose, storage, ontology semantics, or graph display. The current mapping is:

Table 1: VARONA relationship labels across prose, schema, ontology, and interactive graph display.

Natural-language label	Canonical-schema label	Ontology-role label	Interactive relational predicate
Sources	`SOURCE`	`SOURCE`	*is source of*
Policy instruments	`INSTRUMENT`	`INSTRUMENT`	*is used in*
Authorities	`AUTHORITY`	`AUTHORITY`	*has authority over*
Responsible actors	`RESPONSIBLE_ACTOR`	`RESPONSIBLE_ACTOR`	*is responsible for*
Affected actors	`AFFECTED_ACTORS`	`AFFECTED_ACTOR`	*is affected by*
Beneficiaries	`BENEFICIARIES`	`BENEFICIARY`	*benefits from*
Cost bearers	`COST_BEARERS`	`COST_BEARER`	*bears costs of*
Risk bearers	`RISK_BEARERS`	`RISK_BEARER`	*bears risk from*
Locations	`LOCATION`	`LOCATION`	*is location of*

Canonical-schema labels are used in extraction outputs, storage, validation, and network construction. Ontology-role labels identify the semantic function occupied by each entity in a proposition. Natural-language labels are preferred when discussing the concepts substantively. In the interactive visualization, directed entity-to-proposition edges render the same roles as verb-like relational predicates so the graph can be read as a set of relational statements.

Arrowheads make the proposition-mediated direction explicit without changing the underlying ontology labels. Their interpretation remains conditional on the accuracy of the underlying extraction.

Conceptually:

```
Regulator
    |
RESPONSIBLE_ACTOR
    |
    v
[REQUIRE cost protection]
    |
INSTRUMENT
    |
    v
Tariff
```

A richer proposition might have:

```
                        [PROPOSITION]
                             |
          +------------------+------------------+
          |                  |                  |
 RESPONSIBLE_ACTOR        INSTRUMENT          TARGET
          |                  |                  |
          v                  v                  v
     Commission            Tariff        Large-load customer
          |
          +---------------- DISTRIBUTIONAL ----------------+
                             |                             |
                             v                             v
                        BENEFICIARY                   COST_BEARER
                             |                             |
                             v                             v
                        Ratepayers                   Large loads
```

The proposition node preserves the fact that these roles belong to one assertion.

6.2.1 Primary and derived relations

VARONA distinguishes **primary relations** from **derived analytical relations**. Primary relations are the typed entity-to-proposition links extracted from source text.

For example:

```
Ratepayers --BENEFICIARY--> [Policy proposition]
Data centers --COST_BEARER--> [Policy proposition]
Commission --AUTHORITY-----> [Policy proposition]
```

These statements do not, by themselves, assert a direct actor-to-actor edge between ratepayers, data centers, and the commission. An actor-to-actor representation can be derived from their joint participation in a proposition, but that is a second-order analytical projection.

For example, an analyst might define a distributional contrast between a cost bearer and a beneficiary or an institutional relation between a responsible actor and an affected actor. Such projections must be generated by explicit rules and labeled as **derived**, rather than presented as directly observed relations.

This distinction matters because it prevents VARONA from silently converting proposition structure into stronger pairwise claims than the source supports.

The primary network is proposition-mediated; actor-to-actor networks, when useful, are analytical projections of that structure.

6.3 Why qualifications remain properties

Suppose the same proposition also has:

```
status = recommended
modality = should
evidence_basis = stakeholder_judgment
confidence = high
```

Representing each as a separate node would add graph complexity without necessarily improving substantive interpretation. Accordingly, VARONA makes these properties available in the proposition inspector.

6.4 Provenance

Each proposition retains evidence and source-location information.

Provenance provides a path from:

```
network abstraction
        ->
structured proposition
        ->
verbatim evidence
        ->
paragraph/document location
```

This system design allows analysts to verify an interpretation against the source.

7 Analytical Projections

A central VARONA principle is that one underlying representation can support multiple analytical views.

7.1 Complete network

The complete view provides broad access to propositions and entities while preserving the basic graph structure. In the current renderer, front matter (table of contents, letter, and acknowledgments) and bibliographic `reference` records are excluded from the analytical network by default, so this view is complete relative to the configured analytical inclusion rules rather than a literal rendering of every Stage-06 frame.

It is useful for exploration and for identifying clusters of policy activity before narrowing the analysis.

7.2 Institutional structure

The institutional lens emphasizes roles such as:

- Authority.
- Responsible actor.
- Policy instrument.
- Source.

It is designed for questions about governance architecture, jurisdiction, implementation responsibility, and institutional capacity.

7.3 Distributional impacts

The distributional lens emphasizes:

- Affected actors.
- Beneficiaries.
- Cost bearers.
- Risk bearers.

This projection is especially important for policy analysis because a proposal's substantive significance often depends on how consequences are distributed rather than merely on its aggregate effect.

7.4 Stakeholder politics

The political lens emphasizes stakeholder positions and propositions expressing `support`, `conditional support`, `concern`, `dissent`, `opposition`, or `alternatives`.

This projection can support analysis of coalitions, conflict, and contested policy interventions.

7.5 Evidence and certainty

The evidentiary lens focuses on evidence basis and extraction confidence and can be extended to epistemic characteristics such as `observed`, `estimated`, `expected`, `possible`, `conditional`, or `disputed` claims where these are represented.

It is intended to help analysts distinguish policy assertions from the kinds of evidence invoked to support them.

7.6 Substantive lenses

Domain-specific projections can focus on:

- Energy.
- Water.
- Siting.
- Economics.

These lenses allow analysts to move between the overall policy system and particular policy subsystems.

8 Interactive Inspection

The network visualization is not intended to replace detailed reading. It provides an interface for navigating structured information.

8.1 Search and filtering

Users can search proposition content and entities and filter the graph using ontology dimensions, analytical views, and source scope, including:

- Analytical lens and policy attribute domain.
- Proposition type.
- Predicate text through search.
- Controlled action.
- Status.
- Stakeholder position.
- Data-center segment.
- Evidence basis.
- Attribute domain.
- Entity type.
- Semantic relationship role.

8.2 Semantic proposition labels

Internal proposition identifiers are necessary for provenance but are poor human-facing labels. VARONA attempts to construct concise semantic labels from available structured information, preferring a non-generic controlled action with a target, objective, or problem when that normalization is available and reliable, and otherwise using cleaned evidence excerpts.

The source-faithful predicate remains separately available in search and the proposition inspector. The proposition identifier remains available as secondary metadata.

8.3 Proposition inspector

Selecting a proposition exposes a structured inspector organized around analytical meaning rather than a flat database record.

A conceptual inspector is:

```
PROPOSITION

Core proposition
  Type
  Predicate
  Controlled action
  Problem
  Target
  Objective

Policy qualifications
  Status
  Modality
  Condition
  Time
  Threshold

Actors and roles
  Source
  Authority
```

```
  Responsible actor
  Instrument
  Affected actors
  Beneficiaries
  Cost bearers
  Risk bearers

Consequences
  Expected effects
  Adverse effects

Domain attributes
  Energy
  Water
  Siting
  Economics

Evidence and verification
  Evidence basis
  Confidence
  Verbatim evidence

Provenance
  Document
  Paragraph
  Source location
```

Empty or unknown fields can be suppressed to prevent schema completeness from overwhelming substantive information.

8.4 Extraction diagnostics

The inspector can also identify values that warrant review. Examples include:

- A supposed problem that resembles a copied sentence rather than a concise concept.
- A threshold that appears contaminated by citation numbers.
- A proposition with no extracted entity relationships.
- Relationships that exist in the full graph but are hidden under the current projection.

This makes VARONA useful not only for policy analysis but also for iterative evaluation of the LLM-augmented extraction pipeline.

9 What Research Questions Can VARONA Answer?

VARONA's value ultimately depends on the questions its representation makes possible. The system is particularly well suited to questions in which policy content, actors, semantic roles, and evidence must be analyzed jointly.

The questions below describe the analytical potential of the current data model. Whether a particular empirical question can be answered reliably depends on corpus coverage and validated extraction quality.

9.1 Policy agenda and problem structure

VARONA can support questions such as:

- What policy problems are associated with data-center development?
- Which substantive domains dominate the policy agenda?
- Which issues tend to appear together within policy propositions?
- Which policy problems are associated with particular interventions?
- How do energy, water, siting, economic, infrastructure, community, and governance concerns differ in prominence?
- Which data-center segments are associated with particular policy concerns?
- Which problems generate recommendations, implementation actions, or regulatory responses?

These questions move beyond document counts by connecting issue attention to proposition type and action.

9.2 Policy instruments and intervention strategies

The representation can support questions such as:

- What actions are proposed to address particular policy problems?
- Which policy instruments are associated with particular objectives?
- Which interventions rely on regulation, incentives, planning, reporting, coordination, infrastructure investment, or other mechanisms?
- Which interventions are proposed versus already in effect?
- Which actions are associated with particular data-center segments?
- Which policy strategies recur across multiple actors or documents?
- Where are alternative interventions proposed for the same problem?

This can produce an ontology-based map of the policy instrument portfolio.

9.3 Institutional governance

VARONA is designed to answer questions about the architecture of policy responsibility:

- Which institutions claim or possess authority over particular issues?
- Which actors are assigned responsibility for implementation?
- Which institutions are repeatedly associated with particular policy instruments?
- Where is authority concentrated or fragmented?
- Which policy problems require coordination across multiple institutions?
- Where do responsibility and authority appear to be assigned to different actors?
- Which actors function as sources of recommendations but not implementers?
- Which policy domains have clear institutional ownership and which appear institutionally diffuse?

These questions are particularly useful for identifying governance complexity.

9.4 Distributional politics

The explicit representation of affected actors, beneficiaries, cost bearers, and risk bearers enables questions that conventional policy networks often cannot answer directly:

- Who is expected to benefit from a policy?
- Who is expected to bear its costs?
- Who bears risk?
- Which populations are repeatedly described as affected?
- Are the beneficiaries of a policy different from the actors responsible for paying for it?
- Which interventions redistribute costs between data-center operators, utilities, ratepayers, governments, and communities?
- Which groups are simultaneously beneficiaries under some propositions and cost bearers under others?
- Which policy domains generate the greatest distributional conflict?

- Are particular stakeholder positions associated with expected benefits or burdens?

This is one of VARONA's most distinctive analytical capabilities.

9.5 Stakeholder politics and disagreement

Where stakeholder positions are explicitly represented, researchers can ask:

- Which actors support or oppose particular interventions?
- What policy proposals generate reservations or conditional support?
- Which actors propose alternatives?
- Which issues generate the most disagreement?
- Are disagreements primarily about goals, instruments, costs, risks, authority, or evidence?
- Do actors that agree on a problem disagree about the intervention?
- Which policy propositions attract support across different actor types?
- Where do coalitions appear to form?

A sufficiently broad corpus could support analysis of policy coalitions without reducing political positions to simple document co-occurrence.

9.6 Causal and consequential claims

The action and effects vocabularies can support questions such as:

- What is claimed to increase or reduce a policy problem?
- What actions are expected to produce particular outcomes?
- What risks are attributed to data-center development?
- Which interventions are expected to mitigate those risks?
- Which policies are described as producing adverse effects?
- Where do propositions identify tradeoffs between desirable outcomes?
- Are the same interventions associated with both expected benefits and adverse effects?
- Which causal claims are central to particular policy recommendations?

These questions can help reconstruct the implicit theories of intervention contained in policy discourse.

9.7 Evidence, uncertainty, and epistemic structure

VARONA can connect policy claims to their reported evidentiary basis:

- Which propositions rely on empirical facts?
- Which rely on forecasts, estimates, scenarios, legal text, comparative cases, or stakeholder judgment?
- Which policy domains depend most heavily on forecasts?
- Which recommendations are associated with explicit evidence?
- Which high-consequence claims have weak or uncertain evidentiary foundations?
- Where do stakeholders disagree despite invoking similar evidence?
- Where are policy decisions being proposed under explicit uncertainty?
- Which claims should receive priority for human verification?

This can reveal the epistemic architecture of a policy debate, not merely its substantive content.

9.8 Cross-domain policy interactions

Because propositions can carry multiple policy attributes, researchers can investigate relationships between domains:

- How are energy and water concerns connected?

- Which siting interventions have economic consequences?
- Which infrastructure proposals create community impacts?
- Where do grid-reliability objectives conflict with affordability objectives?
- Which interventions address several policy domains simultaneously?
- Which actors bridge otherwise separate policy subsystems?

This can reveal policy interdependencies that are difficult to detect through isolated topic coding.

9.9 Comparative questions

With multiple documents, jurisdictions, organizations, or time periods, VARONA could support comparisons such as:

- How do jurisdictions differ in their data-center policy portfolios?
- Which jurisdictions emphasize regulation versus incentives?
- How do institutional responsibility structures differ?
- Who is treated as a beneficiary or cost bearer across jurisdictions?
- Which policy concerns emerge first and which follow later?
- How do stakeholder positions change over time?
- Does the evidentiary basis of policy claims change as an issue matures?
- Do recommendations eventually become implementation actions or existing policies?
- Which policy innovations diffuse across jurisdictions?

These questions require consistent extraction and appropriate document metadata but do not require a fundamentally different graph model.

9.10 Diagnostic research questions about the extraction system

VARONA can also support methodological questions:

- Which proposition types are extracted most reliably?
- Which ontology fields are systematically underpopulated?
- Which semantic roles are most difficult to distinguish?
- Where does the model confuse descriptive quantities with policy thresholds?
- Which entity classes are most frequently missed or misnormalized?
- Which proposition types produce the greatest disagreement between machine extraction and human coders?
- Does extraction quality vary by document format or policy domain?
- How much graph structure disappears when extraction errors are corrected?

This makes the interface part of the research infrastructure used to validate the system itself.

10 Critical Reports and Analytical Products

VARONA's data can support a family of standardized executive reports. The most useful reports are not simply exports of the network. They reorganize proposition data around policy questions.

10.1 Policy landscape report

Purpose: Provide a high-level account of what the policy corpus contains.

Core components:

- Number of documents, paragraphs, propositions, and normalized entities.
- Propositions by policy domain.
- Propositions by proposition type.

- Actions by frequency.
- Policy status distribution.
- Data-center segments represented.
- Major policy instruments.
- Most frequently represented actor classes.
- Major problem and objective categories.

This report answers:

> What does the policy system contain, and what kinds of policy activity does it represent?

10.2 Institutional responsibility and authority report

Purpose: Map who has authority, who is expected to act, and through which instruments.

Core components:

- Authority by policy domain.
- Responsible actors by action.
- Responsible actor × instrument matrix.
- Policy domain × responsible actor matrix.
- Propositions involving multiple institutional actors.
- Potential authority/responsibility gaps.
- Coordination-intensive propositions.

This report answers:

> Who governs what?

10.3 Distributional impacts report

Purpose: Identify who is expected to benefit, pay, bear risk, or otherwise be affected.

Core components:

- Beneficiaries by policy domain.
- Cost bearers by policy domain.
- Risk bearers by policy domain.
- Affected actors by intervention.
- Beneficiary × cost-bearer relationships.
- Policies with multiple distributional roles.
- Propositions with explicit cost shifting or burden allocation.

This report answers:

> Who gains, who pays, and who bears risk?

10.4 Policy intervention portfolio

Purpose: Describe the instruments and actions used or proposed to address policy problems.

Core components:

- Action × domain.
- Action × proposition type.
- Instrument × objective.
- Problem × action.
- Recommendation × responsible actor.
- Policy status × action.

- Implementation actions associated with recommendations.

This report answers:

> What is government or the policy community proposing to do?

10.5 Stakeholder alignment and conflict report

Purpose: Describe political positions surrounding policy interventions.

Core components:

- Positions by actor.
- Positions by policy domain.
- Positions by action or instrument.
- Support/opposition around individual propositions or intervention families.
- Alternative proposals.
- Contested issues.
- Actor groups with similar position profiles.

This report answers:

> Who agrees, who disagrees, and about what?

10.6 Evidence and uncertainty report

Purpose: Evaluate the evidentiary character of the policy discourse.

Core components:

- Evidence basis by proposition type.
- Evidence basis by domain.
- Forecast-dependent claims.
- Claims based on stakeholder judgment.
- Claims containing uncertainty.
- High-consequence propositions requiring verification.
- Extraction-confidence distributions.
- Propositions with weak or absent evidentiary characterization.

This report answers:

> What kinds of knowledge are policy claims built on?

10.7 Risk and tradeoff report

Purpose: Identify tensions among policy objectives and consequences.

Core components:

- Adverse effects by intervention.
- Risk bearers by action.
- Explicit tradeoff propositions.
- Interventions with both positive and adverse consequences.
- Domain-to-domain tradeoffs.
- Affordability, reliability, water, environmental, and community risks.

This report answers:

> What tradeoffs does the policy system recognize?

10.8 Implementation readiness report

Purpose: Distinguish aspirational recommendations from actionable or implemented policy.

Core components:

- Recommendations by status.
- Implementation actions.
- Responsible actors.
- Authority relationships.
- Policy instruments.
- Conditions and thresholds.
- Implementation timelines.
- Propositions missing an explicit responsible actor or authority.

This report can help distinguish:

> What is being discussed from what can actually be implemented?

10.9 Extraction quality and ontology coverage report

Purpose: Support human validation of the generated dataset.

Core components:

- Propositions with and without entity links.
- Ontology-field completion by proposition type.
- Semantic-role coverage.
- Suspicious threshold values.
- Unusually long or sentence-like structured fields.
- Low-confidence frames.
- Entity normalization anomalies.
- Human-adjudication queue.

This is a methodological report rather than a substantive policy report, but it is critical whenever VARONA data are used for research.

11 Critical Visualizations

VARONA's network is only one visualization. The underlying proposition table supports several other visual forms that may be more appropriate for particular questions.

11.1 Interactive proposition network

Best for: exploration, tracing relationships, and inspecting individual propositions.

Recommended visual grammar:

- Circles = entities.
- Diamonds = propositions.
- Proposition fill = proposition type.
- Typed edges = semantic participant roles.
- Lens-specific secondary visual encoding.
- Selection = structured inspector plus evidence.

The network should prioritize interpretability over displaying every available field.

11.2 Policy-domain × action heatmap

Rows represent policy domains and columns represent controlled actions.

Cell values can represent:

- Proposition counts.
- Proportions within domain.
- Validated weighted counts where appropriate.

This visualization quickly shows the intervention vocabulary associated with each policy domain.

11.3 Problem → intervention Sankey or alluvial diagram

A flow visualization can connect:

```
Problem -> Action -> Instrument -> Objective
```

or, where data quality supports it:

```
Problem -> Recommendation -> Responsible actor -> Expected effect
```

This is useful for showing how the policy system translates diagnosed problems into interventions.

It should be used selectively: flows with large numbers of categories can become unreadable.

11.4 Institutional responsibility matrix

A matrix can place responsible actors on one axis and policy domains or actions on the other.

This can reveal:

- Concentration of responsibility.
- Institutional specialization.
- Shared responsibility.
- Domains with no clearly identified implementer.

A second matrix can perform the same function for authority.

11.5 Distributional flow diagram

A particularly valuable VARONA visualization can connect:

```
Policy action
    ->
Affected actor
    ->
Beneficiary / Cost bearer / Risk bearer
```

Alternative versions can directly compare beneficiaries and cost bearers by intervention.

This visualization operationalizes one of the most distinctive parts of the ontology.

11.6 Beneficiary–burden matrix

Rows can represent beneficiary classes and columns can represent cost-bearer classes, with cells showing the number or proportion of propositions connecting each pair.

This provides a compact representation of distributive politics.

11.7 Stakeholder-position matrix

Rows represent actors or actor classes and columns represent interventions, domains, or policy instruments.

Cells encode:

- Support.
- Conditional support.
- Reservation.
- Dissent.
- Opposition.
- Alternative.

This is often easier to interpret than a political network when the goal is comparison rather than exploration.

11.8 Evidence-basis profile

A stacked bar chart or heatmap can show evidence basis by:

- Domain.
- Proposition type.
- Action.
- Actor class.
- Document.

This makes it possible to compare, for example, policy domains dominated by quantitative estimates with domains dominated by stakeholder judgment.

11.9 Evidence × consequence matrix

One axis represents evidence basis and the other expected/adverse consequence categories.

This can help identify consequential claims resting primarily on forecasts, scenarios, stakeholder judgments, or empirical observations.

11.10 Policy-status pipeline

Where status data are reliable, a flow or ordered bar visualization can distinguish:

```
exploratory
    ->
proposed
    ->
recommended
    ->
planned
    ->
underway
    ->
in effect / completed
```

This should not be interpreted as a temporal transition unless longitudinal evidence establishes that the same policy actually moved through these stages.

11.11 Policy-attribute small multiples

Energy, water, siting, and economic attributes can be visualized as separate small-multiple charts rather than forcing all domains into one figure.

Examples include:

- Energy source × action.
- Grid issue × proposition type.
- Water-use issue × intervention.
- Siting issue × responsible actor.
- Economic issue × beneficiary.

11.12 Graph coverage dashboard

A methodological dashboard should report:

- Total propositions.
- Total entities.
- Semantic-role edges.
- Propositions connected to entities.
- Propositions with no entity links.
- Role coverage.
- Field completion.
- Extraction confidence.
- Flagged values.

This dashboard is critical because analysts should know the representational coverage of the graph before drawing conclusions from its topology.

11.13 Proposition evidence cards

For communication with policy practitioners, individual propositions can be rendered as compact evidence cards:

```
RECOMMENDATION

Action
Require ...

Responsible actor
...

Affected actors
...

Expected effect
...

Evidence
"...verbatim passage..."

Source
Document / paragraph
```

A collection of such cards can be more useful than a network for decision-makers who need traceable findings rather than graph exploration.

12 A Core VARONA Reporting Suite

Although many analyses are possible, a practical implementation should prioritize a small set of reproducible products.

The most critical suite is:

1. **Policy Landscape**
   - Domains, proposition types, actions, instruments, actors.
2. **Institutional Map**
   - Authority, responsibility, coordination, instruments.
3. **Distributional Map**
   - Affected actors, beneficiaries, cost bearers, risk bearers.
4. **Stakeholder Politics**
   - Support, opposition, reservations, alternatives, contested interventions.
5. **Evidence and Uncertainty**
   - Evidence basis, forecasts, uncertainty, confidence, verification priorities.
6. **Intervention and Implementation Portfolio**
   - Recommendations, implementation actions, responsible actors, status, conditions, thresholds.
7. **Ontology and Extraction Quality**
   - Coverage, missing roles, suspicious fields, confidence, adjudication priorities.

Together these reports move from:

```
WHAT IS BEING DISCUSSED?
        |
        v
WHAT IS BEING PROPOSED?
        |
        v
WHO HAS TO ACT?
        |
        v
WHO BENEFITS / PAYS / BEARS RISK?
        |
        v
WHO SUPPORTS OR OPPOSES IT?
        |
        v
WHAT EVIDENCE SUPPORTS IT?
        |
        v
HOW RELIABLE IS THE STRUCTURED REPRESENTATION?
```

That final question is deliberately part of the reporting architecture rather than an afterthought.

13 Development Application: Arizona Data-Center Policy

VARONA is being developed using policy documents concerning data-center development in Arizona, including the Arizona Energy Promise Taskforce material processed through the current pipeline.

The case is useful for system development because data-center policy intersects multiple substantive domains and cross-cutting policy dimensions:

- Electricity demand.
- Generation.
- Transmission.
- Grid reliability.
- Utility rates and cost allocation.
- Water use.
- Cooling.
- Siting.
- Permitting.
- Workforce.
- Economic development.
- Local community impacts.
- Institutional authority.
- Implementation.

It also contains a heterogeneous mixture of recommendations, descriptive claims, forecasts, implementation proposals, stakeholder statements, quantitative information, and institutional descriptions.

This heterogeneity makes the corpus a demanding test for proposition extraction.

13.1 Why the application is currently a development case

The Arizona application should not yet be treated as a finalized empirical dataset. Iterative development runs have identified several classes of extraction error, including:

- Citation numbers incorrectly interpreted as quantitative thresholds.
- Long source spans placed in fields intended for concise problem concepts.
- Non-propositional participant lists incorrectly converted into propositions.
- Participant entities assigned indiscriminately across multiple distributional roles.
- Beneficial outcomes misclassified as adverse effects.
- Lexical cues incorrectly triggering proposition types.
- Recommendation headings at risk of being filtered too aggressively by proposition-existence rules.

These errors are informative for system development. They demonstrate why provenance and human inspection are necessary components of an ontology-guided LLM workflow.

Several refinements have already shown that targeted prompt constraints can substantially reduce specific failure modes. Nevertheless, empirical claims about the Arizona policy system should be based on a validated extraction protocol rather than intermediate development output.

13.2 Demonstration rather than substantive inference

At the current stage, the Arizona case can appropriately demonstrate:

- How propositions are represented.
- How entities occupy semantic roles.
- How analytical lenses alter the graph projection.
- How evidence is recovered for inspection.
- How extraction anomalies become visible.

- How ontology coverage can be measured.

It should not yet be used to make definitive claims about the prevalence of particular policy relationships without validation.

14 Validation and Evaluation

A proposition-centered knowledge network is only as useful as the validity of the information used to construct it.

VARONA therefore requires evaluation at several levels.

14.1 Proposition detection

The first question is whether the system correctly distinguishes propositions from non-propositional material.

Evaluation should measure false positives arising from:

- Headings.
- Rosters.
- Affiliation lists.
- Table fragments.
- Citations.
- Labels.
- Isolated noun phrases.

It should also measure false negatives, particularly substantive propositions embedded in headings or numbered recommendation statements.

14.2 Proposition-type classification

Human coders can evaluate whether extracted propositions receive the appropriate type.

Particular attention should be paid to boundaries such as:

- Substantive claim versus evidence claim.
- Recommendation versus implementation action.
- Stakeholder position versus substantive problem description.
- `WATER_CLAIM` versus `GRID_CLAIM` for cooling-related electricity demand.
- Existing policy versus recommendation.

14.3 Predicate and controlled-action extraction

The source-faithful predicate and the optional controlled action should be evaluated separately. Human coding should assess whether the predicate preserves the operative assertion and whether a populated controlled action is a faithful analytical normalization. Evaluation should also verify that descriptive propositions are allowed to retain a meaningful predicate with `action = null` rather than being forced into an ill-fitting controlled category.

14.4 Entity normalization

Evaluation should examine:

- Entity detection recall.
- Entity-class accuracy.
- Canonicalization accuracy.
- Erroneous merges.
- Duplicate canonical entities.
- Treatment of generic actors and affected populations.

14.5 Semantic-role assignment

Semantic roles require separate validation because identifying the correct participants does not guarantee assigning them correctly.

Key distinctions include:

- Affected actor versus beneficiary.
- Beneficiary versus cost bearer.
- Cost bearer versus risk bearer.
- Authority versus responsible actor.
- Source versus responsible actor.

This is especially important for distributional analysis.

14.6 Proposition qualifications

Fields such as target, objective, threshold, expected effect, adverse effect, status, and modality should be evaluated for both presence and semantic correctness.

For example, threshold validation should distinguish genuine policy triggers from ordinary descriptive quantities.

14.7 Evidence grounding

Every frame should be checked to determine whether its verbatim evidence actually supports the extracted proposition.

This is a critical safeguard against plausible but unsupported structure.

14.8 Human adjudication

A validation workflow should use sampled propositions and trained human coders to evaluate extraction performance rather than relying solely on anecdotal inspection (Gilardi, Alizadeh, and Kubli 2023; Törnberg 2024; Wang et al. 2024; Kim et al. 2024).

Evaluation metrics can include precision, recall, F1, class-specific performance, and agreement measures appropriate to each field.

The exact validation design and empirical performance statistics remain to be established for the finalized protocol and should not be inferred from development runs.

15 VARONA as a Human-in-the-Loop Monitoring Tool

VARONA's interactive network has a methodological role beyond presenting completed results.

A conventional extraction pipeline often has the following structure:

```
documents -> model -> structured dataset -> analysis
```

The risk is that errors become less visible after conversion into rows, fields, nodes, and edges. Structured output can create an unwarranted appearance of precision.

VARONA adds an inspection loop:

```
documents
    ->
ontology-guided extraction
```

```
    ->
structured propositions
    ->
VARONA
    ->
human inspection
    ->
diagnosis
    ->
prompt / ontology / validation refinement
    ->
re-extraction
```

This workflow makes visualization part of data-quality assessment.

For example, exposing a `threshold` property in the proposition inspector can reveal that citation numbers have been interpreted as thresholds. A distributional lens can reveal that the same actors have been copied into beneficiary and cost-bearer roles. A proposition label can expose malformed targets. A disconnected proposition can reveal either genuinely absent entity roles or extraction failure.

These are not merely interface defects. They are observable consequences of the information-generation process.

VARONA therefore supports an iterative model of computational research in which extraction, representation, visualization, and validation inform one another (Tsaneva et al. 2025; Wang et al. 2024).

16 Discussion

16.1 From text retrieval to policy structure

The main conceptual contribution of VARONA is a shift from retrieving policy-relevant text to representing policy-relevant structure.

Keyword search can retrieve a paragraph about utility rates. VARONA aims to represent that paragraph as one or more propositions specifying:

- What is claimed or proposed.
- Which actors participate.
- What roles they occupy.
- What intervention is involved.
- What consequences are asserted.
- What evidence is invoked.
- Where the proposition came from.

This enables questions that require relations rather than isolated categories.

16.2 Beyond actor networks

Policy-network analysis often focuses on actors and their relationships. VARONA can complement that tradition by representing policy propositions as mediating objects.

Instead of assuming a direct relationship:

```
Actor A <-> Actor B
```

the system can represent:

```
Actor A
   |
   | SUPPORTS / SOURCE / RESPONSIBLE_ACTOR
   v
Policy proposition
   ^
   | AFFECTED_ACTOR / BENEFICIARY / COST_BEARER
   |
Actor B
```

This preserves the substantive reason the actors are related.

16.3 Distributional analysis as a first-class capability

Explicit beneficiary, cost-bearer, risk-bearer, and affected-actor roles are particularly valuable.

Policy debates frequently revolve around distribution:

- Who receives infrastructure.
- Who pays for it.
- Who bears environmental or financial risk.
- Who receives economic benefits.
- Who is protected by a regulation.

A generic actor network obscures these distinctions. VARONA makes them queryable, subject to extraction validity.

16.4 Epistemic policy analysis

Policy disagreements can involve different values or interests, but they can also involve different understandings of evidence and uncertainty.

By retaining evidence basis and epistemic qualifications, VARONA creates the possibility of studying not only:

> What does an actor claim?

but also:

> What kind of evidence is presented as supporting the claim?

This is especially relevant in emerging policy domains where forecasts, scenarios, uncertain resource estimates, and rapidly changing technologies influence decision-making.

16.5 Analytical lenses as projections, not separate datasets

The lens architecture is important conceptually. Institutional, distributional, political, and evidentiary views should not be understood as unrelated networks.

They are projections of the same proposition-centered representation.

This allows an analyst to move from a policy recommendation to:

- Its institutional implementation structure.
- Its distributional consequences.
- Its political support or opposition.
- Its evidentiary basis.

without losing the common proposition that connects these dimensions.

17 Limitations

VARONA has several important limitations.

17.1 Dependence on extraction quality

The graph does not eliminate language-model errors. It can make them more visible, but substantive analysis still depends on validated extraction.

17.2 Ontology dependence

Ontology-guided extraction necessarily privileges concepts specified in advance. Relevant information outside the ontology may be missed or poorly represented.

The ontology should therefore be treated as an explicit analytical model, not a neutral description of reality.

17.3 Missing does not equal absent

A missing beneficiary does not establish that no beneficiary exists. It may mean:

- No beneficiary was stated.
- The field was not applicable.
- The model failed to extract it.

This limits interpretations based on missingness.

17.4 Proposition granularity

One paragraph can contain several propositions, and one sentence can contain nested or conditional propositions. Deciding how finely to split them is an analytical choice.

Excessive splitting can destroy context; insufficient splitting can conflate distinct claims.

17.5 Semantic-role ambiguity

Some actors legitimately occupy multiple roles. A utility might be both responsible for implementing an intervention and affected by it. Anti-duplication rules therefore cannot simply prohibit all multiple-role assignments.

The requirement is semantic justification for each role.

17.6 Visualization limits

Large proposition networks can become visually dense. No graph layout can make every relationship simultaneously legible.

VARONA addresses this through filtering, lenses, semantic labels, and inspectors, but tabular and statistical visualizations remain necessary complements.

17.7 Corpus dependence

The network represents the documents supplied to the pipeline, not the entire policy system.

Claims about actor importance, policy prevalence, or institutional centrality therefore require a defensible corpus design.

17.8 Temporal inference

Policy status categories should not automatically be interpreted as stages in a temporal process. Demonstrating policy change requires longitudinal evidence linking propositions across time.

18 Future Development

Several extensions follow naturally from the current architecture.

18.1 Multi-document and multi-jurisdiction corpora

Scaling VARONA across jurisdictions would enable comparative analysis of policy portfolios, institutional structures, distributional politics, and policy diffusion.

18.2 Temporal policy networks

With reliable dates and proposition alignment, the system could represent how:

- Problems emerge.
- Recommendations appear.
- Stakeholder positions change.
- Policies move toward implementation.
- Evidence accumulates.
- Coalitions evolve.

18.3 Proposition alignment

Semantically similar propositions could be grouped into proposition families while retaining source-specific instances.

This would help distinguish:

```
same policy idea repeated across documents
```

from:

```
independent policy propositions
```

18.4 Contradiction and disagreement detection

A future layer could identify propositions that make incompatible claims about:

- Causal relationships.
- Costs.
- Benefits.
- Authority.
- Feasibility.
- Evidence.

Such a feature should remain evidence-grounded rather than relying on unconstrained model judgments.

18.5 Human adjudication interface

VARONA could evolve from an inspection tool into an adjudication environment where analysts:

- Accept.
- Reject.
- Correct.
- Relabel.
- Annotate.

individual proposition fields and semantic relationships.

The corrected records could become validated research data and evaluation sets.

18.6 Structured query and export

The proposition-centered representation could be exported to graph databases and other structured formats, enabling queries such as:

```
Find all recommendations
about electricity cost allocation
where ratepayers are beneficiaries
and data-center operators are cost bearers.
```

or:

```
Find all water-related implementation actions
assigned to local governments
that cite quantitative evidence.
```

18.7 Automated analytical reporting

Once extraction has been validated, VARONA data could feed reproducible report generators producing standardized policy landscape, institutional, distributional, political, and evidence reports.

These reports should always retain links back to proposition evidence.

19 Implications for Policy Research

VARONA suggests a broader approach to generative AI in policy research.

LLMs are often used as summarizers: documents enter the model and prose summaries emerge. That can be useful, but it produces outputs that are difficult to aggregate, compare, query, and validate systematically.

Ontology-guided information extraction takes a different approach. The model is used as one component of a research pipeline that converts text into explicitly defined analytical objects.

The intended workflow is not:

```
document -> LLM -> answer
```

but:

```
document
   ->
bounded textual units
   ->
ontology-guided extraction
   ->
normalized entities
   ->
grounded propositions
   ->
```

```
structured knowledge representation
   ->
human inspection
   ->
validated analysis
```

In this model, the ontology constrains what is generated, provenance constrains what can be claimed, and human inspection constrains how model output is used.

This architecture may be particularly useful in policy domains characterized by large document collections, heterogeneous actors, technical complexity, and rapidly evolving debates.

20 Conclusion

VARONA is a proposition-centered approach to transforming unstructured policy documents into inspectable policy knowledge networks.

Its central premise is simple: policy documents contain more than topics and entities. They contain structured assertions about what problems exist, what actions should occur, which institutions have authority, who is responsible for implementation, which instruments are used, who benefits, who bears costs and risks, what consequences are expected, what positions stakeholders take, and what evidence supports those claims.

Representing these distinctions requires a richer architecture than entity co-occurrence alone.

VARONA therefore places policy propositions at the center of the graph and connects canonical entities to those propositions through typed semantic roles.

Additional qualifications remain available as proposition properties. Analytical lenses expose institutional, distributional, political, evidentiary, and substantive dimensions without requiring every ontology element to appear simultaneously on the interface. Verbatim evidence and provenance keep the resulting abstraction connected to its documentary source, facilitating human inspection.

This representation supports analysis across several dimensions without requiring separate datasets for each one. Users can move from questions about institutional responsibility or distributional consequences to stakeholder disagreement, evidence, or cross-domain relationships while working from the same proposition-centered data. The same structure can also feed analytical reports and visualizations for executive decision-making.

At the same time, VARONA does not remove the need for validation. Its development has demonstrated that ontology-constrained extraction can still produce semantically plausible but incorrect structure. The interactive network is therefore designed not only to communicate findings but also to expose the information-generation process to scrutiny.

The resulting approach treats generative AI neither as an autonomous policy analyst nor merely as a text summarizer. Instead, it uses language models within an ontology-guided, evidence-preserving, human-inspectable computational workflow (Kommineni, König-Ries, and Samuel 2024; Tsaneva et al. 2025).

The long-term objective is a research infrastructure in which analysts can move systematically between policy documents, propositions, actors, relationships, evidence, networks, and reproducible analytical products without losing sight of the text on which every structured claim ultimately depends.

Appendix

Appendix A: Core Analytical Objects

Table 2: Core analytical objects represented or retained by VARONA.

Object	Function
Document	Original policy source
Paragraph	Bounded provenance unit
Entity mention	Textual reference to an actor or policy object
Canonical entity	Normalized representation of an entity
Proposition	Structured policy assertion
Proposition type	Kind of assertion being made
Predicate	Source-faithful operative assertion
Action	Optional controlled analytical normalization
Semantic role	Entity's function within a proposition
Qualification	Status, modality, condition, threshold, etc.
Effect	Expected or adverse consequence
Evidence basis	Character of evidence supporting a proposition
Verbatim evidence	Source text grounding the extraction
Confidence	Extraction-level confidence indicator

Appendix B: Core Analytical Lenses

Table 3: Core analytical lenses supported by the proposition-centered representation.

Lens	Primary question	Principal information
Complete	What does the policy knowledge network contain?	All supported proposition/entity relationships
Institutional	Who governs and implements?	Authority, responsible actors, instruments, sources
Distributional	Who gains, pays, or bears risk?	Affected actors, beneficiaries, cost bearers, risk bearers
Stakeholder politics	Who supports or opposes what?	Positions, alternatives, contested interventions
Evidence & certainty	What supports the claims?	Evidence basis, uncertainty, confidence
Energy	What is happening in energy/grid policy?	Energy propositions and attributes
Water	What is happening in water/cooling policy?	Water propositions and attributes
Siting	What is happening in siting/permitting?	Siting propositions and attributes
Economics	What is happening in economic/workforce policy?	Economic propositions and attributes

Appendix C: Recommended Reporting Products

Table 4: Recommended reporting products supported by VARONA's structured data.

Product	Primary audience	Core question
Policy Landscape	Analysts and decision-makers	What is the policy system discussing?
Institutional Map	Government and implementation teams	Who has authority and responsibility?
Distributional Map	Policy analysts and communities	Who benefits, pays, and bears risk?
Intervention Portfolio	Policy designers	What interventions are proposed or underway?
Stakeholder Politics	Public-affairs and governance analysts	Where is agreement or conflict?
Evidence & Uncertainty	Researchers and decision-makers	What evidence supports policy claims?
Risk & Tradeoff Report	Strategic decision-makers	What adverse consequences and tradeoffs are recognized?
Implementation Readiness	Implementers	Which proposals have actors, authority, instruments, and conditions?
Extraction Quality	Researchers	How trustworthy and complete is the structured dataset?

Appendix D: Recommended Visualization Portfolio

Table 5: Recommended visualization portfolio for VARONA outputs.

Visualization	Best analytical use
Interactive proposition network	Exploration and evidence inspection
Domain × action heatmap	Policy strategy comparison
Problem → intervention alluvial	Problem-to-response structure
Institutional responsibility matrix	Governance structure
Distributional flow diagram	Benefits, costs, and risks
Beneficiary–burden matrix	Distributional politics
Stakeholder-position matrix	Alignment and conflict
Evidence-basis profile	Epistemic structure
Evidence × consequence matrix	Evidentiary support for consequential claims
Policy-status pipeline	Policy maturity/status, with appropriate temporal caution
Domain small multiples	Detailed substantive comparison
Graph coverage dashboard	Extraction and representation diagnostics
Proposition evidence cards	Traceable policy communication

Ashwin, Julian, Aditya Chhabra, and Vijayendra Rao. 2023. "Using Large Language Models for Qualitative Analysis Can Introduce Serious Bias." *Sociological Methods & Research*, 00491241251338246.

Barrasa, Jesús, and Jim Webber. 2023. *Building Knowledge Graphs.* " O'Reilly Media, Inc.".

Bisbee, James, and Arthur Spirling. 2026. "What to Do When Humans Are No Longer the Gold Standard: Large Language Models, State of the Art and Robustness."

Brach, William, Kristián Košt'ál, and Michal Ries. 2025. "The Effectiveness of Large Language Models in Transforming Unstructured Text to Standardized Formats." *IEEE Access*.

Eisenstein, Jacob. 2019. *Introduction to Natural Language Processing*. MIT Press.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks." *Proceedings of the National Academy of Sciences* 120 (30): e2305016120.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, et al. 2021. "Knowledge Graphs." *ACM Computing Surveys (Csur)* 54 (4): 1–37.

Keet, C Maria. 2018. *An Introduction to Ontology Engineering*. University of Cape Town.

Kim, Hannah, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. "Meganno+: A Human-Llm Collaborative Annotation System." In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 168–76.

Kommineni, Vamsi Krishna, Birgitta König-Ries, and Sheeba Samuel. 2024. "From Human Experts to Machines: An LLM Supported Approach to Ontology and Knowledge Graph Construction." *arXiv Preprint arXiv:2403.08345*.

Liu, Kang, Yubo Chen, Jian Liu, Xinyu Zuo, and Jun Zhao. 2020. "Extracting Events and Their Relations from Texts: A Survey on Recent Research Progress and Challenges." *AI Open* 1: 22–39.

Sewerin, Sebastian, Lynn H Kaack, Joel Küttel, Fride Sigurdsson, Onerva Martikainen, Alisha Esshaki, and Fabian Hafner. 2023. "Towards Understanding Policy Design Through Text-as-Data Approaches: The Policy Design Annotations (POLIANNA) Dataset." *Scientific Data* 10 (1): 896.

Törnberg, Petter. 2024. "Best Practices for Text Annotation with Large Language Models." *arXiv Preprint arXiv:2402.05129*.

Tsaneva, Stefani, Danilo Dessì, Francesco Osborne, and Marta Sabou. 2025. "Knowledge Graph Validation by Integrating LLMs and Human-in-the-Loop." *Information Processing & Management* 62 (5): 104145.

Tudorache, Tania. 2020. "Ontology Engineering: Current State, Challenges, and Future Directions." *Semantic Web* 11 (1): 125–38.

Wang, Xinru, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. "Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels." In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–21.

Xie, Zhengnan, Alice Saebom Kwak, Enfa George, Laura W. Dozal, Hoang Van, Moriba K. Jah, Roberto Furfaro, and Peter Alexander Jansen. 2022. "Extracting Space Situational Awareness Events from News Text." In *International Conference on Language Resources and Evaluation*. https://api.semanticscholar.org/CorpusID:246015662.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics* 50 (1): 237–91.

Zufall, Elise, and Tyler A Scott. 2024. "Syntactic Measurement of Governance Networks from Textual Data, with Application to Water Management Plans." *Policy Studies Journal* 52 (4): 941–54.