TrustFinder

Recommendations for a Community-Based System for Finding Trusted Sources and Evaluating Claims

Lead Designer: R.J. Cordes

Contributors:
Scott David J.D., L.L.M.
Daniel Friedman, PhD

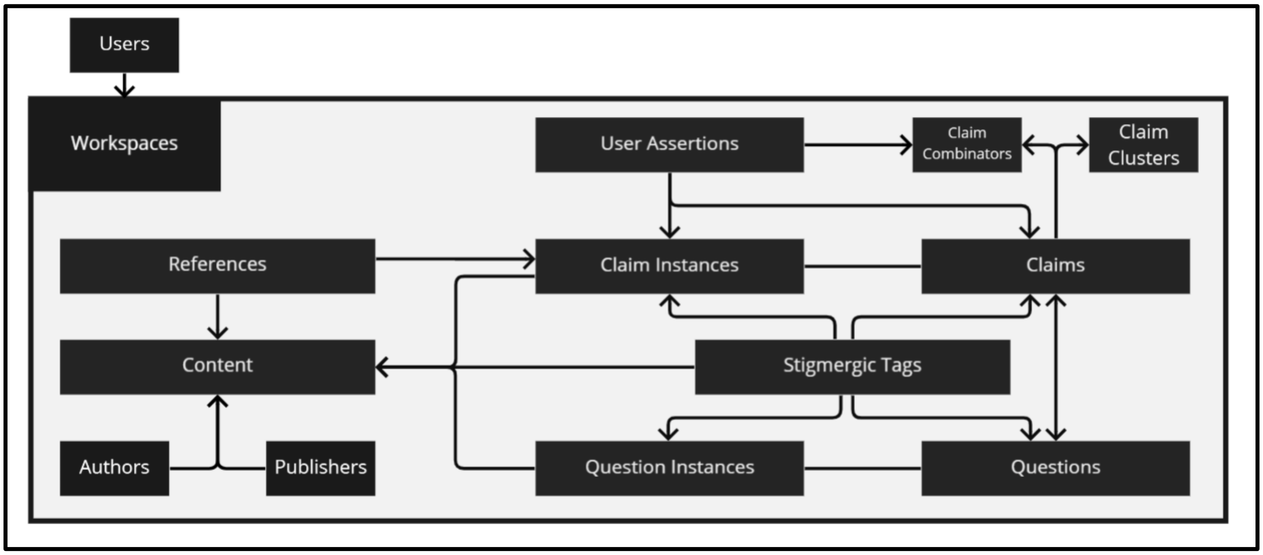
Consultants:
Mridula Mascarenhas, PhD

August 25, 2022
v. 1.0

Executive Summary

There is a broadly recognized need for better situational awareness within the information environment. Each year, millions of articles, books, documents, and datasets are published. Amidst this flood of information, even those with significant experience and expertise in the knowledge economy are struggling to evaluate and vet claims. This document builds on the feedback of dozens of experts across myriad fields submitted to the University of Washington Applied Physics Lab's Verified Information Exchange Environments Program, to present recommendations for a sociotechnical system, "TrustFinder", for collaborative management of the information supply chain. TrustFinder implements controls and standards, web and document annotation affordances, argument representation frameworks, and crowdsourcing design principles in order to harness the work of global research communities. The ultimate goal of TrustFinder is to structure the information environment to such an extent that it enables users to find trusted sources of information and rapidly assess concepts and claims.

System Overview



TrustFinder Environment Primary Components

TrustFinder has one primary category of actors, “**Users**”. **Users** scope the global information environment (i.e. the internet) by creating, sharing, and adding other users to “**Workspaces**”, which represent “information commons” intended to facilitate projects related to sensemaking (e.g., a research paper, studying, exploration of a topic). **Users** within **Workspaces** use web and document annotation affordances (i.e., the ability to “mark-up”, “highlight”, take notes at the edge of a document or webpage, or to otherwise enrich content) in order to structure the information environment. **Users** and **Workspaces** can further assign “trust scores” representing expectations of the quality and intents of specific authors and publishers, as well as of the assertions and annotation contributions by other **Users** and **Workspaces**. With such enrichment tools, **Users** structure the claims and concepts they encounter in order to make the information environment more navigable and searchable, reducing future redundant work for themselves and others related to evaluation and vetting of claims and allowing for evaluation and mapping of the information supply chain (i.e., where claims originate and where they have spread).

In the TrustFinder environment, a **Workspace** can be populated with different classes of interconnected informational structures, each contributing to enrichment of the rhetorical landscape. Below are the 10 primary classes of informational elements.

Content. Information sources such as a book, paper, article, or video. Content provides a container for **References**, user-added metadata, **Claim Instances**, and **Question Instances**.

References. Relationships between **Content**, such as direct citations, are stored as references. These reference objects can be used to map the connections between **Content** and b between **Claim Instances**.

Claim. A statement about the world can be represented as a Claim. Claims exist outside the context of any specific **Content**, can be represented using various phrasings, and can be connected to other objects. A **Workspace** can be prompted or initiated using a Claim, as a basis to help scope related work (i.e., *this work* is related to the investigation of *this Claim*). The Claim's most important feature is its ability to be connected to other Claims through **Claim Combinators** and **Claim Clusters**.

Question. Explicit or implicit Questions are represented by an informational structure which can be connected to both **Claims** and other Questions**.** Similar to **Claims**, they can exist outside the context of any specific **Content**, can be represented using various phrasings, can prompt or initiate a **Workspace**, and be connected to other objects. Questions have an important relationship with **Claims** as **Claims** can both be responses to, or prompt, Questions.

Claim Instance. As opposed to **Claims**, which exist outside the context of any specific **Content**, Claim Instance objects represent the instantiation, or appearance, of a particular **Claim** within a specific area of a piece of **Content** (i.e., within a particular sentence). Claim Instances can be connected to the appearances of its **Claim** within other pieces of **Content** through **References** (i.e., where there is a direct citation related to the appearance of the Claim Instance within the **Content**).

Question Instance. Similar to the **Claim Instance**, Question Instances are simply instantiations, or appearances, of a particular **Question** within a specific area of **Content**.

Claim Cluster. Claim Clusters are a simple container for **Claims** that are related in terms of their relationship to some other object (i.e., this set of **Claims**, if all true, support this other **Claim**).

Claim Combinator. Claim Combinators are containers for describing the relationship between a **Claim** or a **Claim Cluster**, and a **Claim**, **Claim Cluster**, or another Claim Combinator. Claim Combinators are categorized as (i) supports, (ii) refutes, (iii) generalizes, (iv) modifies, and (v) relates to.

User Assertion. In addition to collecting **Claims** and marking **Claim Instances**, **Users** can also make their own assertions *about* **Claims** and **Claim Instances**. User Assertions are essentially a special form of **Claim Combinator**, on which they are attaching their name. User Assertions attached to **Claims** will appear within the workspace when **Users** access the **Claim**, as well as when they access instances of that claim (**Claim Instance**), allowing for contextualization of particular claims. User Assertions attached to **Claim Instances** will only appear on that particular **Claim Instance**, allowing for nuanced warnings or endorsements (e.g., if you wanted to find support for this **Claim**, this particular piece of content may not be the place to cite it from, as it is not a strong argument or works from faulty data).

Stigmergic Tag. Stigmergic Tags are a combination of predefined and **User-**defined tags used to further assist in querying and navigating **Workspaces**. Stigmergic Tags provide users with highly structured methods for communicating requests, directing attention, providing feedback, and marking the presence of key concepts or entities. Stigmergic Tags can be connected to nearly all other informational structures within the TrustFinder environment, including other Stigmergic Tags.

System Purpose

The primary purpose of the sociotechnical system, “TrustFinder”, is to facilitate collaborative structuring of the information environment, enabling users to find trusted sources of information, which in turn enables them to rapidly assess concepts and claims. The secondary purposes include:

- providing infrastructure and data for the future of reference management systems,

- mapping and understanding the "supply-chain" of claims, and
- "capturing" the value of discourse and disagreement.

Scope

This document intends to provide recommendations for the key components of the TrustFinder environment and for their structure and relationships from the perspective of knowledge management and behavioral modification in the context of crowdsourcing solutions, as well as to offer (i) relevant background information regarding the basis of these recommendations and (ii) a discussion of the potential implications of their implementation. It does not provide (i) exhaustive recommendations for user experience or presentation, or (ii) detailed recommendations or technical requirements for data structure or security assurances. Names for components within these recommendations should be adapted to optimize user experience and onboarding. **A developed TrustFinder system may differ substantially from recommendations given technical constraints or opportunities.**

Structure of this Document

This document consists of (i) a Systems Definition section concerned with the components of the TrustFinder system, separated into 5 segments: (a) Agents and Workspaces, (b) Media, (c) Claims, (d) Questions, and (e) Reputation; (ii) an Implications section, which discusses the potential implications of explicit and implicit mechanisms within the recommended system; and (iii) a Background section, which provides a synthesis of theory and frameworks used to inform design. Within the Systems Definition section, explicit mentions of system components are bolded outside of their respective sections for reference purposes. Component attributes, related interfaces, and other objects are bolded and/or italicized for clarity where necessary.

Definitions and Word Usage

"Combinator" is used within this document to describe an empty interface that allows a set of objects which do not necessarily share common methods or attributes to be used in fields which establish complex relationships between said objects. Borrowed and adapted from library organization design patterns within the Haskell programming community, wherein "combinators" are used to combine values of a given type in various ways to create more complex, and context-rich instances of that type.

"Decorator" is used within this document to describe an empty interface used in order to allow a set of objects which do not necessarily share common methods or attributes to be used in a field without modifying the behavior or structure of that object.

“Genuine Presence Testing” describes the set of security assurances which use biometrics, computer vision, and geographic data related approaches to authenticate the presence of a particular person using a device.

“Interface” is used within this document to refer to general “programming interfaces” unspecific to any language, i.e., (i) an object which enables polymorphism, (ii) an object which represents a contract fulfilled by the ability to perform some function or deliver some attribute, or (iii) a vehicle for the inclusion of multiple classes of object within a field which requires type assertion (i.e., a Decorator).

“TrustFinder Environment” is used to describe the space of engagement with the common TrustFinder infrastructure generally, through workspaces or otherwise.

System Definition

Agents and Workspaces

User

User refers to users of TrustFinder, individuals who are seeking to enrich web and document content and collect and evaluate claims. Users must engage with security assurances (e.g., genuine presence testing) in order to register and engage with certain aspects of the system (e.g., **User Assertions**).

Invitation Tree

Users can invite others to TrustFinder. Each **User** invited, and each invited by those invitees, up to 6 degrees of separation, are included within the inviter's Invitation Tree with their respective degree of separation (see Figure 1, degrees of separation). Invitation trees are not visible to other **Users**, and are used primarily to provide foundation for network-related impact scoring. It is recommended that in the future, there are methods devised to allow users to share the credit of invitation of new members and that invitation trees related to specific workspaces (i.e., tracking invitations to workspaces, as opposed to the platform as a whole), are implemented.

Real World Credentials

Users can attach real world credentials, such as higher education degrees and professional certifications to their account.

Pseudonyms

Users can create multiple Pseudonyms (i.e., usernames) for use within the TrustFinder environment. **Users** may selectively disclose which credentials, if any, and what aspects of those credentials to attach to Pseudonyms (e.g., “a Master's degree in computer science” as opposed to “a Master's degree from *this* university”). Pseudonyms may be used to engage with any activity within the TrustFinder environment with the exception of **User Assertions**, which must be tied directly to the **User's** account.

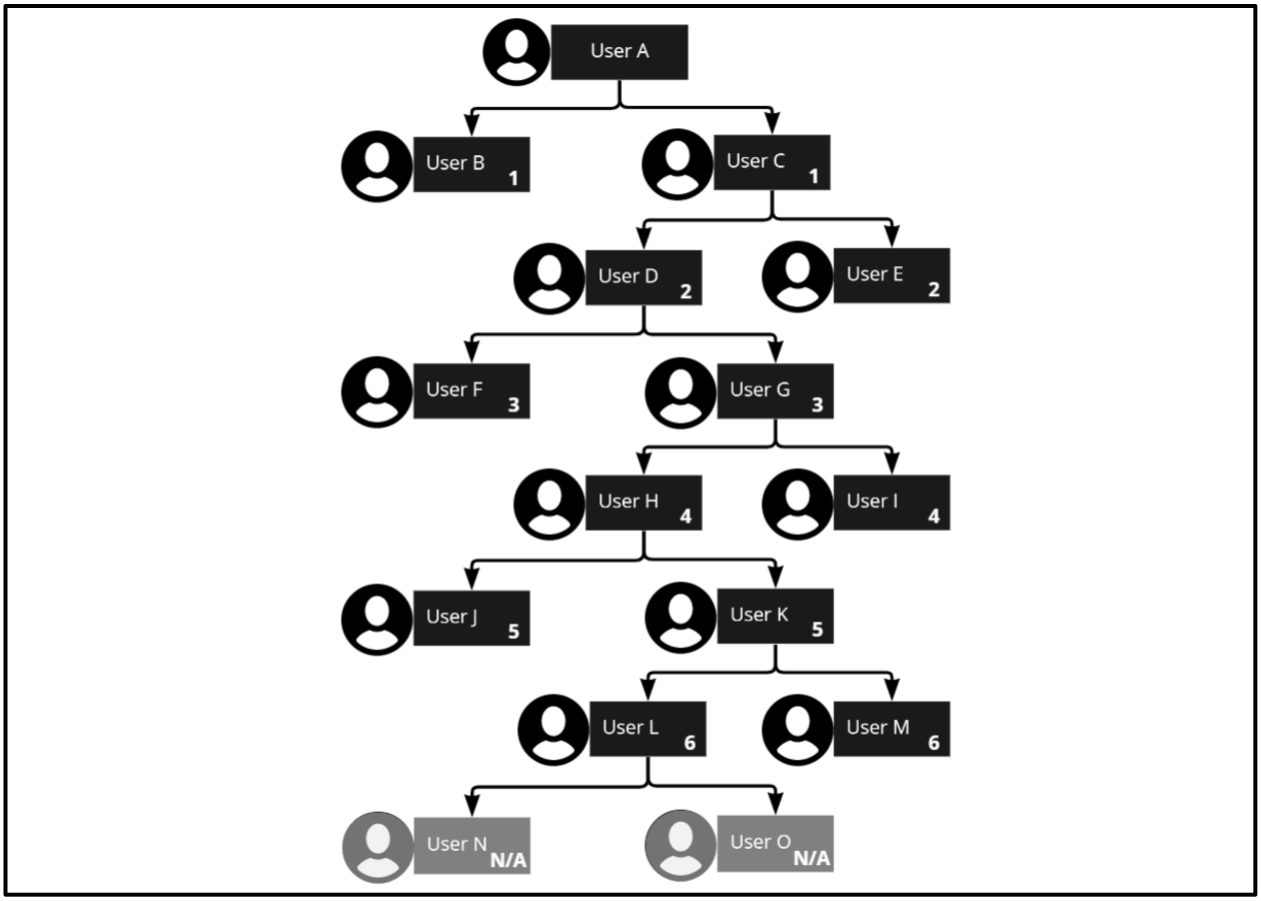


Figure 1. User Invitation Tree

Workspace

Workspaces are the basis for engagement within the TrustFinder environment. **User's** may create and be invited to multiple Workspaces. Workspaces represent projects related to sensemaking (e.g., a research paper, studying, exploration of a topic), and are used as containers for objects relevant to that work.

- Workspaces may be instantiated using a **Claim** or **Question** (e.g., where research on a particular question is the driving motive behind intended work) and can be populated with ***Workspace Objects*** based on the presence of certain **Stigmergic Tags** within those objects, as well as other conditions (e.g., time period, object type).
- Workspaces make use of ***Clearinghouses*** in order to manage the dynamic import and export of digital goods (i.e., ***Workspace Objects***).
- Workspaces may be given their own sets of ***Entity Tag Types***, ***Custom Tag Types***, **Contribution Trust Scores**, and **Assertion Trust Scores**.

- Workspaces have two classes of **User** within, *Administrators* and *Members*. Administrators have permission to manage high level aspects of the Workspace, including setting ***Clearinghouse*** import and export conditions, ***Entity Tag Types**, **Custom Tag Types**,* and the Workspace's **Contribution Trust Scores** and **Assertion Trust Scores**. It is recommended that, at the outset, role and permissions related governance are kept as simple as practicable, while allowing for opportunities to adapt and related features in response to need and interest.

Workspace Object

Workspace Object is a decorator for the following objects: Authors, Publishers, Artifacts, URLs, Content, References, Claims, Claim Instances, Claim Combinators, Claim Clusters, Questions, Question Instances, Question Combinators, User Assertions, and Stigmergic Tags.

Clearinghouse

The Clearinghouse represents the import or export channel for *Workspace Objects* between the **Workspace** and another **Workspace** or set of **Workspaces**. It contains conditional statements for managing the dynamic (i.e., active or ongoing) import and export of ***Workspace Objects***, and a *Buffer*. The *Buffer* is used where **Workspace** administrators opt to approve items individually before they are added to the local **Workspace** environment or before they are available for export to external **Workspace**. Clearinghouses are directional, with export-oriented Clearinghouses making digital goods "available" based on conditional statements to the **Workspaces** specified, allowing those **Workspaces** to create respective import Clearinghouses in response; and with import-oriented Clearinghouses acting as "listening posts" waiting for exports to be made available.

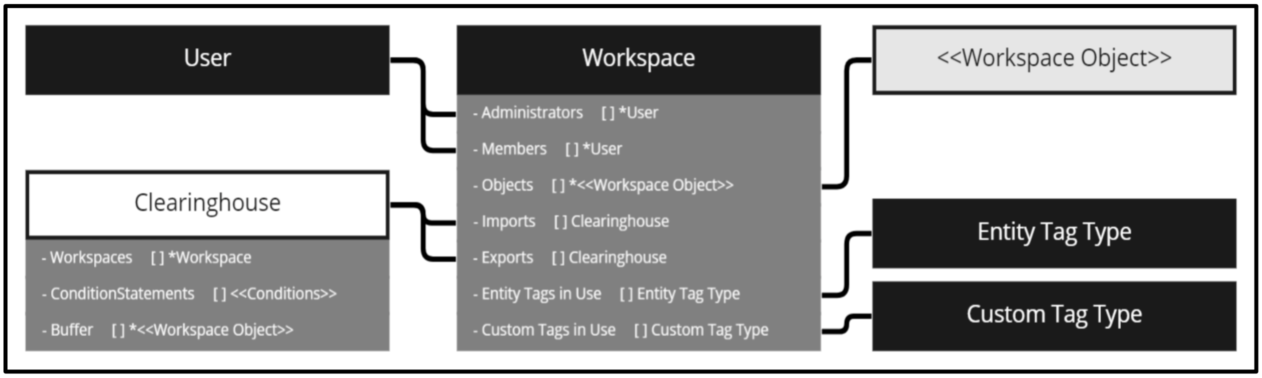

User
Workspace
- Administrators [ ] *User
- Members [ ] *User
- Objects [ ] *<<Workspace Object>>
- Imports [ ] Clearinghouse
- Exports [ ] Clearinghouse
- Entity Tags in Use [ ] Entity Tag Type
- Custom Tags in Use [ ] Custom Tag Type
<<Workspace Object>>
Clearinghouse
- Workspaces [ ] *Workspace
- ConditionStatements [ ] <<Conditions>>
- Buffer [ ] *<<Workspace Object>>
Entity Tag Type
Custom Tag Type

Figure 2. Workspace relationships

Media

Author

An Author object is used to represent authors responsible for **Content**. Authors can be assigned to **Content** (for attribution). **Users** and **Workspaces** may assign Authors an **Assertion Trust Score**. Authors may be given additional attributes over time, such as funding sources, affiliations, and academic credentials and professional certifications.

Publisher

A Publisher object is used to represent the publisher responsible for **Content**. Publishers can be created and assigned to **Content** (for attribution). **Users** and **Workspaces** may assign Publishers an **Assertion Trust Score**. Publishers may be given additional attributes over time, such as funding sources and parent organizations.

Artifact

Artifacts are an object used to represent a stable container for **Content**, such as a PDF or JPG. Artifacts can be linked together as “near duplicates”, where the contents and identifiers are identical, but the resulting hash of the contents are not as a result of file type, resolution, or other adaptations. It is recommended that a combination of Artifact data and data from linked **Content** objects be used as a basis for defining annotation presentation when viewing the Artifact.

URL

The URL object is used to represent unstable, potentially dynamic, web-hosted containers for **Content**. The URL object is recommended to be paired with the use of link-rot and content change detection approaches in order to alert **Workspace** members to potential **Content** changes. It is recommended that a combination of URL data and data from linked **Content** objects be used as a basis for defining annotation presentation when viewing the URL.

Content

The Content object is used to represent units of referenceable information. As such, it might represent an entire book, a book chapter, an area under a subheading, a segment of an image, an entry in a glossary, etc. The Content object can point to other Content objects contained within (e.g., a chapter in a book, or a subheading in a chapter, a figure in a subheading), can point to other variants (e.g., a translated version, a republishing), and be found across multiple **Artifacts**. Content is expected to be assigned an **Author**, **Publisher**, and *Date of Release,* and can contain **Claim Instances**, **Question Instances**, and **References**.

Reference

The Reference object is a decorator for the following objects: **Direct References** and **Implied References**.

Direct Reference

The Direct Reference object is used to mark labeled, explicit references within **Content** to external **Content**. A Direct Reference must be labeled with a "Type", such as "*in-text reference*", "*footnote*", "*endnote*", or "*in-text citation*", indicating the style through which it presents the reference.

Implied Reference

The Implied Reference object is used to mark what the **User** believes to be an implied reference within **Content** to external **Content**. An Implied Reference must be labeled with the contributing **User's** measure of *Certainty* [0-1] about the implication (i.e., "how likely is it that the **Author** was referencing the external **Content**?").

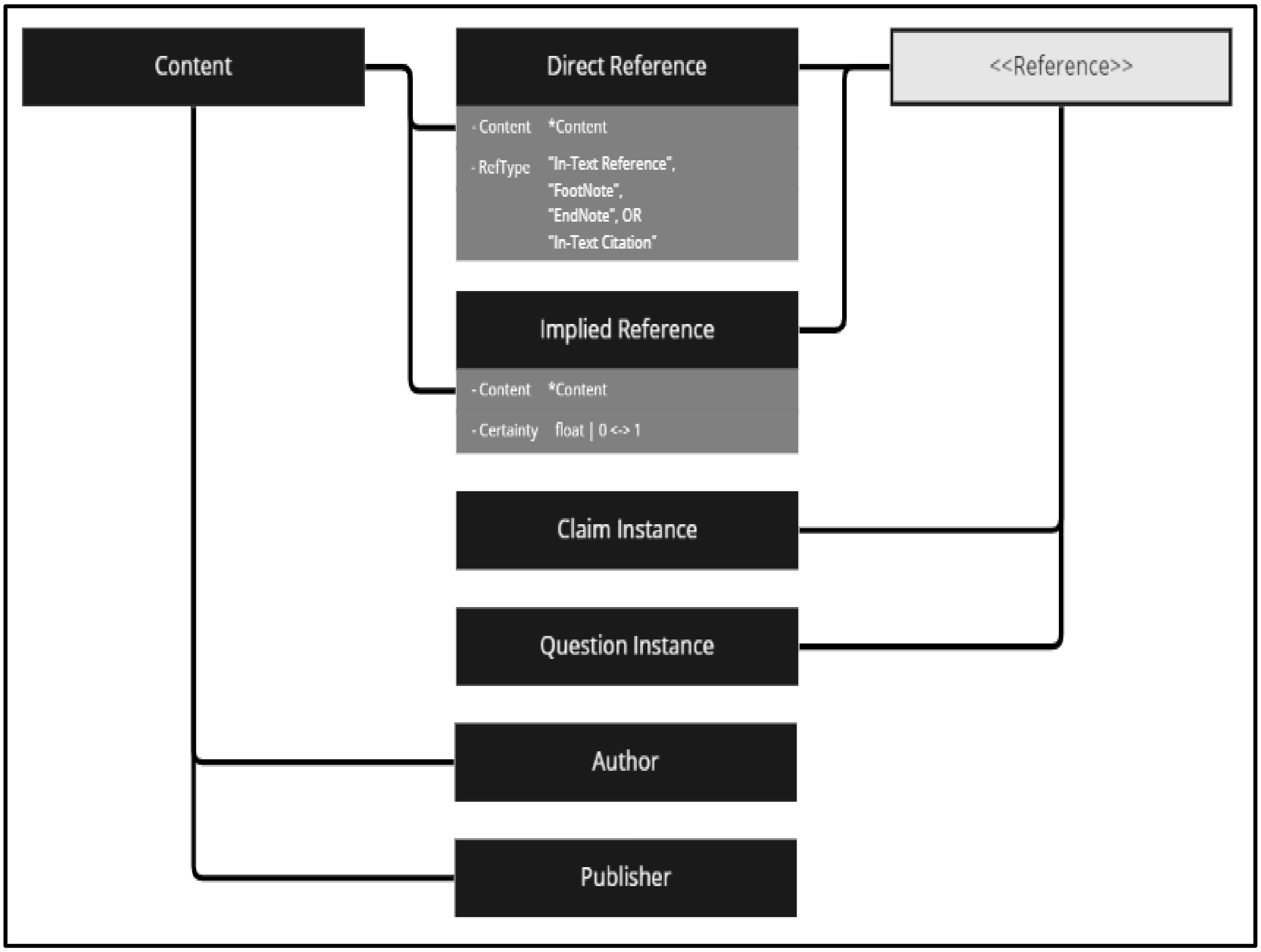


Figure 3. Reference relationships

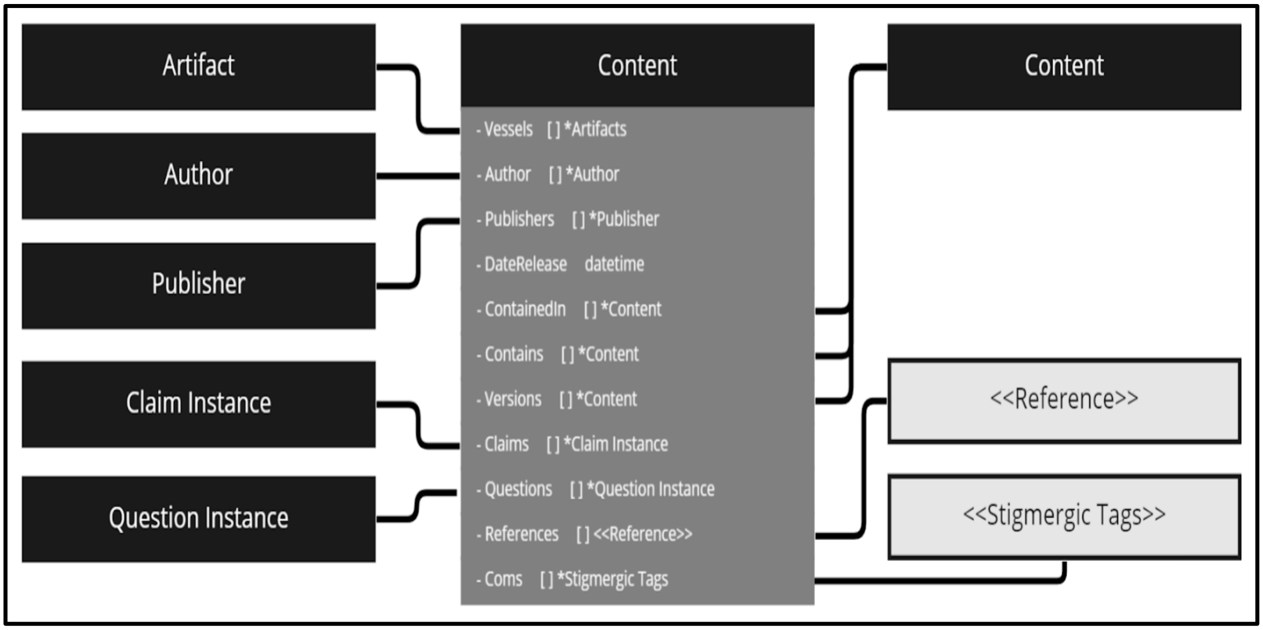


Figure 4. Content relationships

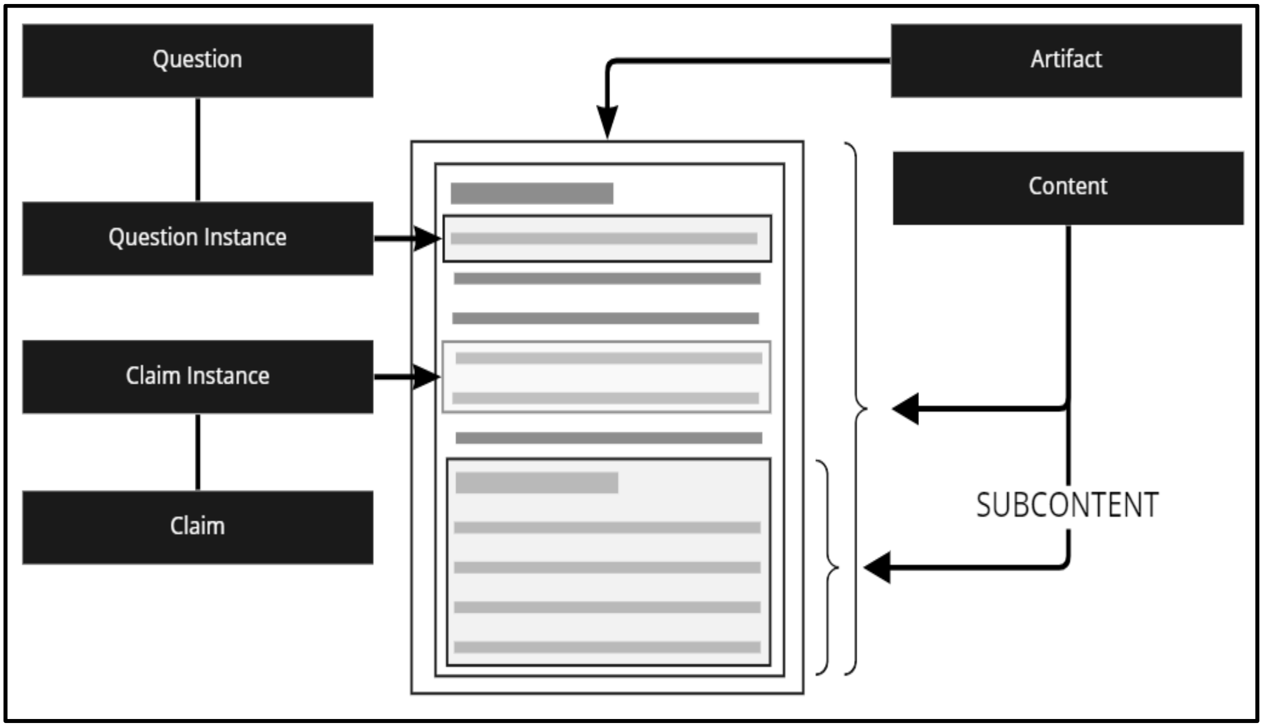


Figure 5. Graphical representation of subcontent and annotation within content

Claims

Claim

A Claim is an object which contains a phrase and variants on that phrase which express a “claim” or assertion. Claims also contain a field for *Counterclaims,* or Claims which assert the exact opposite of the subject Claim (e.g., “x is an integer” and “x is not an integer”).

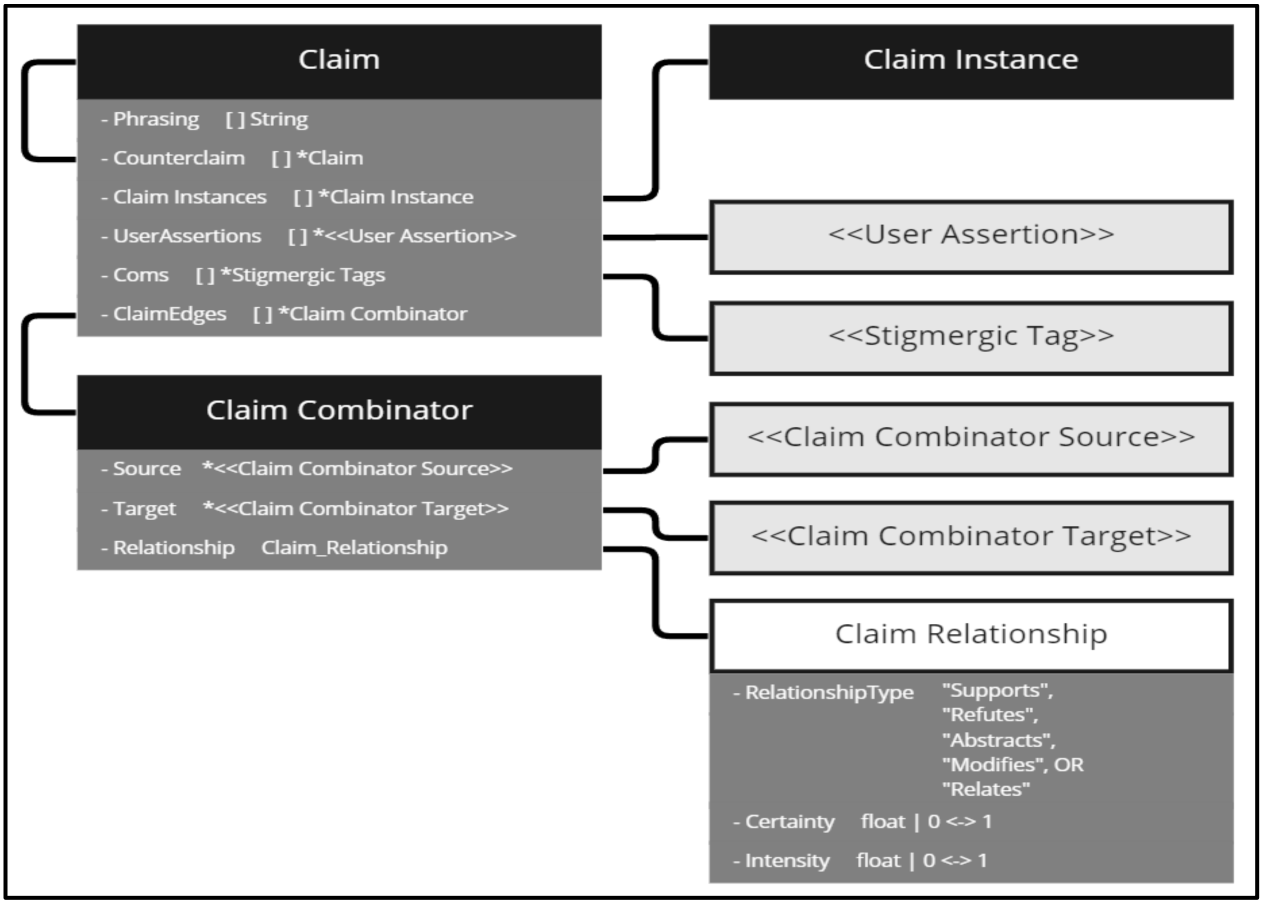

Claim
- Phrasing [ ] String
- Counterclaim [ ] *Claim
- Claim Instances [ ] *Claim Instance
- UserAssertions [ ] *<<User Assertion>>
- Coms [ ] *Stigmergic Tags
- ClaimEdges [ ] *Claim Combinator
Claim Instance
<<User Assertion>>
<<Stigmergic Tag>>
Claim Combinator
- Source *<<Claim Combinator Source>>
- Target *<<Claim Combinator Target>>
- Relationship Claim_Relationship
<<Claim Combinator Source>>
<<Claim Combinator Target>>
Claim Relationship
- RelationshipType "Supports", "Refutes", "Abstracts", "Modifies", OR "Relates"
- Certainty float | 0 <-> 1
- Intensity float | 0 <-> 1

Figure 6. Claim and Claim Combinator relationships

Claim Combinator

The Claim Combinator object is the basis for forming directional relationships, or edges, between **Claims, Claim Clusters**, and other Claim Combinators. Claim Combinators are composed of a **Claim Combinator Source, Claim Combinator Target**, and ***Claim Relationship***.

Claim Relationship

A Claim Relationship adds context to a **Claim Combinator**. It is composed of a *Relationship Type*, which describes the relationship between the **Claim**

Combinator Source and the **Claim Combinator Target** contained within the **Claim Combinator**; and a 2 dimensional vector containing the contributing **User's** (i) *Intensity* [0-1] rating (e.g., how much does the **Claim Combinator Source** *support* the **Claim Combinator Target**) and (ii) their *Certainty* [0-1] rating (i.e., how certain the **User** is of this relationship between the **Claim Combinator Source** and the **Claim Combinator Target**). There are 5 available *Relationship Types,* and while each is directional - there is an implied bidirectionality (e.g., where Object A *supports* Object B, Object B *is supported by* Object A).

- **Supports | Is Supported By.** Where the **Claim Combinator Source** *supports* the **Claim Combinator Target** (e.g., "*x* is an integer less than 2" -> *supports* -> "*x* is equal to 1").
- **Refutes | Is Refuted By.** Where the **Claim Combinator Source** *refutes* the **Claim Combinator Target** (e.g., "*x* is an integer less than 2"-> *refutes* -> "x is equal to 3").
- **Generalizes | Specifies.** Where the **Claim Combinator Source** *generalizes* the **Claim Combinator Target**, in that it is a generalized version of the same claim (e.g., "*x* is a symbol" -> *generalizes* -> "x is a mathematical variable").
- **Modifies | Is Modified By.** Where the **Claim Combinator Source** *modifies* the **Claim Combinator Target**, in that it is a modified version of a similar claim, in that it has added conditions, refinement, or mutations (e.g., "x is a mathematical variable in the context of *this* equation" -> *modifies* -> "x is variable").
- **Relates To | Relates To.** Where the **Claim Combinator Source** *relates to* the **Claim Combinator Target**, in that it is similar, communicates something about the other, or shares a context (e.g., "*x is an integer*" -> *relates to* -> "*y* is an integer").

Claim Combinator Source

A Claim Combinator Source is a decorator for the following objects: **Claims** and **Claim Clusters**.

Claim Combinator Target

A Claim Combinator Target is a decorator for the following objects: **Claims**, **Claim Clusters**, and **Claim Combinators**.

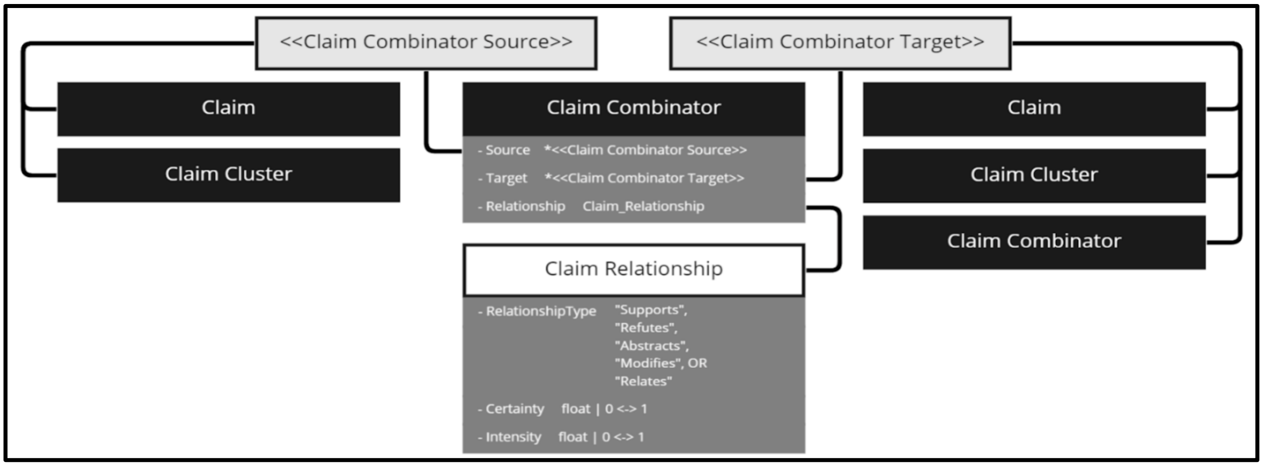

<<Claim Combinator Source>>
<<Claim Combinator Target>>
Claim
Claim Cluster
Claim Combinator
- Source *<<Claim Combinator Source>>
- Target *<<Claim Combinator Target>>
- Relationship Claim_Relationship
Claim Relationship
- RelationshipType "Supports", "Refutes", "Abstracts", "Modifies", OR "Relates"
- Certainty float | 0 <-> 1
- Intensity float | 0 <-> 1
Claim
Claim Cluster
Claim Combinator

Figure 7. Claim Combinator relationships

Claim Instance

A Claim Instance is an object representing the annotation of the presence of a **Claim** within a particular piece of **Content**. A Claim Instance must be labeled with the contributing **User's** measure of *Certainty* [0-1] about the Claim Instance (i.e., "how likely is it that this **Claim** is what the **Author** is discussing or asserting?").

- A **User** may mark a Claim Instance as *Asserted True, Asserted False,* or *Discussed*, in order to indicate whether the **Author** of the **Content** is asserting the relevant **Claim** is True or False, or simply discussing it, respectively.
- A **User** may mark a Claim Instance as *Explicit* or *Implicit*, in order to indicate that the **Author** of the **Content** is discussing the underlying **Claim** directly, or if the **Claim** is latent or implied in the **Content**.

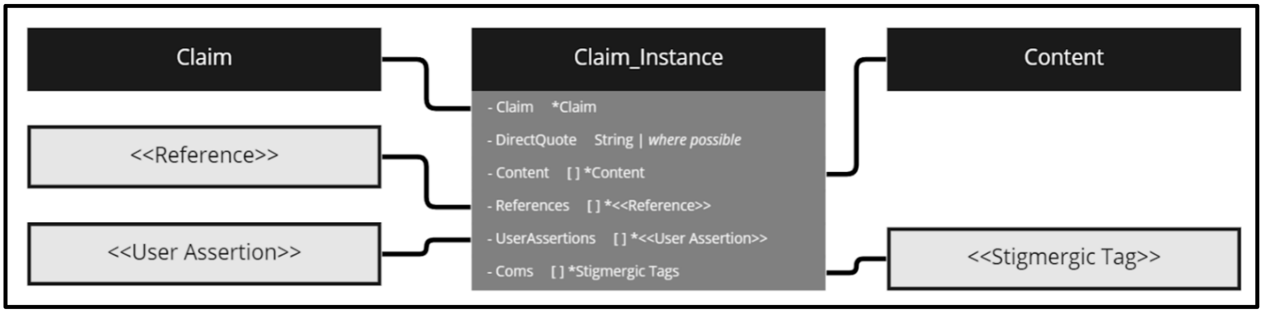

Claim
<<Reference>>
<<User Assertion>>
Claim_Instance
- Claim *Claim
- DirectQuote String | where possible
- Content [ ] *Content
- References [ ] *<<Reference>>
- UserAssertions [ ] *<<User Assertion>>
- Coms [ ] *Stigmergic Tags
Content
<<Stigmergic Tag>>

Figure 8. Claim Instance relationships

Claim Cluster

A Claim Cluster is a container for a set of **Claims** which are grouped together for the purpose of conjecture, context, or collation (e.g., ["*x* is an integer", "*x* is a positive number", "*x* is a number less than 2", "*x* is a number greater than 0"] -> ***supports*** -> "x is equal to 1").

Questions

Question

A Question is an object which contains a phrase and variants on that phrase which express a “question”, *Prompts* (***Question Combinators*** which might inspire or beg the question), and *Responses* (***Question Combinators*** which might be answers or responses to the question).

Question Combinator

A Question Combinator is a decorator for the following objects: **Claims**, **Claim Clusters**, and **Questions**.

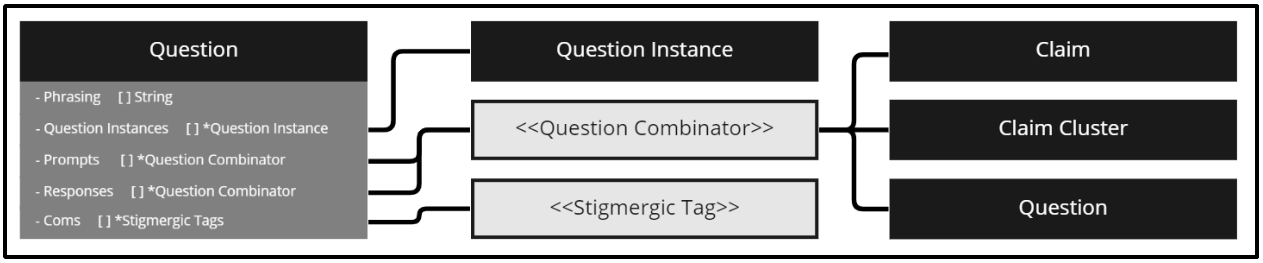

Question
- Phrasing [ ] String
- Question Instances [ ] *Question Instance
- Prompts [ ] *Question Combinator
- Responses [ ] *Question Combinator
- Coms [ ] *Stigmergic Tags
Question Instance
<<Question Combinator>>
<<Stigmergic Tag>>
Claim
Claim Cluster
Question

Figure 9. Question relationships

Question Instance

A Question Instance is an object representing the annotation of the presence of a **Question** within a particular piece of **Content**. A Question Instance must be labeled with the contributing **User’s** measure of *Certainty* [0-1] about the Question Instance (i.e., “how likely is it that this **Question** is what the **Author** is discussing or asking?”).

- A **User** may mark a Question Instance as *Explicit* or *Implicit*, in order to indicate whether the **Author** of the **Content** is discussing the underlying **Question** directly, or if the **Question** is latent or implied in the **Content**.

User Communications

User Assertion

A User Assertion is an interface representing a **User's** personal assertion about the truth or falsity of a particular **Claim** or **Claim Instance** in the form of an ***Endorsement*** or ***Warning*** object. Unlike other annotation affordances, which may be contributed to **Workspaces** by a **User's** chosen pseudonym, it must be attached to the **User's** account. A User Assertion may be attached to either a **Claim** or a **Claim Instance**, creating an option to offer either a Global or Local assertion - as a User Assertion attached to a **Claim Instance** will only be available when interacting with that particular instance of the claim in some **Content**, whereas a User Assertion attached to a **Claim** would be available both during interactions with that **Claim** object, but also during interactions with any of its instantiations (i.e., **Claim Instances**). A **User Assertion** must be accompanied by a plain text explanation, and a 2-dimensional vector containing the contributing **User's** (i) *Intensity* [0-1] rating (e.g., "How untrue or true is this claim?", and (ii) *Certainty* [0-1] rating (e.g., i.e., how certain the User is of this evaluation). It may also be accompanied by ***User Assertion Support*** objects, such as additional **Claims**. It is recommended that **Users** be required to engage with identity assurance tests (e.g., Genuine Presence Testing) in order to post User Assertions.

User Assertion Target

A User Assertion Target is a decorator for the following objects: **Claims** and **Claim Instances**.

User Assertion Support

A User Assertion Support is a decorator for the following objects: **Claims**, **Claim Clusters**, and **References**.

Warning

User Assertions which are intended to warn others of the contents of a **Claim** or **Claim Instance** (e.g., "this claim may be false", "this claim is certainly false and is likely made in bad faith")

Endorsement

User Assertions which are intended to endorse the contents of a **Claim** or **Claim Instance** (e.g., "this claim may be true", "this claim is certainly true, and could only be refuted in bad faith").

Stigmergic Tag

A Stigmergic Tag is a decorator for the following objects and interfaces: **Requests**, **Rallies**, **Remarks**, **Entity Tag Instances**, and **Custom Tag Instances**, each of which is intended to structure **User** communications at scale and can be attached to nearly any other TrustFinder object (exception being **Users** and **Workspaces**) including other Stigmergic Tags**.**

Request

A Request is an interface for **Stigmergic Tags** which ask or "ping" other **Users** within a **Workspace** to engage in a specific action. Requests can be suggested to be resolved by those who respond, and may be resolved by **Workspace** administrators or the original contributor of the Request. A Request must be accompanied by the contributing **User's** *Intensity* [0-1] rating (i.e., "how urgent or important is it that this request be responded to?").

- **Skeptical.** A request for clarification about an object or topic from a position of skepticism (i.e., uncertainty with an interest in evaluation).
- **Curious.** A request for clarification about an object or topic from a position of curiosity (i.e., uncertainty with an interest in exploration).
- **Search.** A request for more information about an object or topic which may be already known or more easily searchable by other members of the **Workspace** (e.g., "are there other papers on this specific phenomena mentioned here?").
- **Catalog.** A request specifically intended to prompt the annotation or cataloging of information found by someone more capable (e.g., "please annotate this potential **Claim Instance**").
- **Custom Request. Workspaces** can implement local, specific **Request** tags to meet their own needs.

Rally

A Rally is a special **Stigmergic Tag** which adds to the *Intensity* rating of other **Stigmergic Tags** in order to help direct attention within a **Workspace** and reduce the likelihood of duplicates or simply directs attention to a particular object. A Rally must be accompanied by the contributing **User's** *Intensity* [0-1] rating (i.e., "how urgent or important is it that others see this?").

Remark

A Remark is a **Stigmergic Tag** which is used to add plain text for miscellaneous comments. A Remark must be accompanied by the contributing **User's** *Intensity* [0-1] rating (i.e., "how urgent or important is it that others see this?").

Entity Tag

An Entity Tag is a tag which indicates the presence of a reference (not to be confused with **References**) in **Content** to a specific entity, such as a concept, idea, person, place, or thing.

- **Entity Tag Type.** An Entity Tag Type is a container for the schema and details of a Custom Tag (e.g., attributes and respective values of the particular Entity Tag, related entities, parent and child Entity Tags, and aliases).

Custom Tag

A Custom Tag is an interface for **Stigmergic Tags** named and implemented by **Workspaces** for local use. It acts as a compensating control for where no other tag structure is adequate for what the **Workspace** needs to represent or mark.

- **Custom Tag Type.** Custom Tag Type is a container for the schema and details of a Custom Tag.

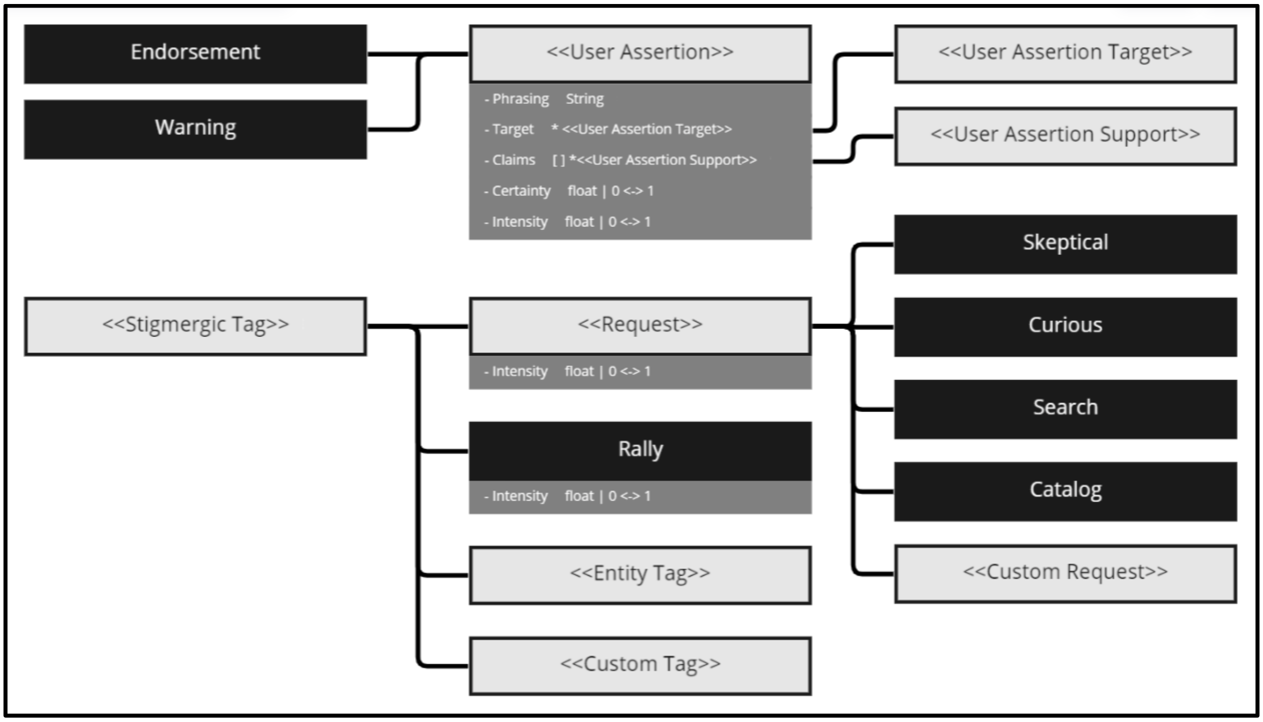

Endorsement
Warning
<<User Assertion>>
- Phrasing String
- Target * <<User Assertion Target>>
- Claims [ ] *<<User Assertion Support>>
- Certainty float | 0 <-> 1
- Intensity float | 0 <-> 1
<<User Assertion Target>>
<<User Assertion Support>>
Skeptical
<<Stigmergic Tag>>
<<Request>>
- Intensity float | 0 <-> 1
Curious
Search
Rally
- Intensity float | 0 <-> 1
Catalog
<<Entity Tag>>
<<Custom Request>>
<<Custom Tag>>

Figure 11. User Communications relationships

Reputation

There are five separate base types of Reputation in the TrustFinder system: **CQ Annotation Score**, **Mapping Impact Score**, **Network Impact Score**, **Contribution Trust Score**, and **Assertion Trust Score**. Each is a relatively simple metric representing a signal of trust based on past interactions which can be used within the TrustFinder environment or by third parties in order to generate other, optional forms of reputation calculation metrics to **Users**. Nearly all are defined exclusively through set construction and calculation of cardinality, with the only exception being **Network Impact Score**, which uses set construction in combination with a standard decay function.

CQ Annotation Score

The Contribution Quality (CQ) Annotation Score is a simple metric intended to represent the volume of a particular **User's** direct contributions to identifying claims and questions found in **Content** (i.e., **Claim Instances, Question instances**), within the context of a **Workspace** or a collection of **Workspaces**. Relevant objects include (i) **Claim Instances** and **Question Instances** where the **User** was the initial contributor (e.g., the discoverer of a given claim), and (ii) the **Claim Instances** and **Question Instances** which were contributed within or imported to a given **Workspace** or collection of **Workspaces**. The **CQ Annotation Score** (CQAS) is defined as the cardinality of the set of **Claim Instances** and **Question Instances** formed from the intersection of the set of **Claim Instances** and **Question Instances** by a given **User** (C), with the union of the sets of **Claim Instances** and **Question Instances** associated with a given collection of **Workspaces** (W).

$$\mathrm{CQAS}(\mathrm{C}\,\{\,\}, \{\mathrm{W}_{\mathrm{i}}\{\,\}, \ldots\}) = |\mathrm{C} \cap (\mathrm{W}_{\mathrm{i}} \cup \mathrm{W}_{\mathrm{i}+1} \cup \mathrm{W}_{\mathrm{i}+\ldots})|$$

- Every **Workspace** has a dynamically calculated **CQ Annotation Score** for each **User** which has contributed relevant objects within, either as members or as a result of imports from other **Workspaces**. This collection of scores includes scores for **Users** who are not members of that **Workspace** but have contributions present as a result of imports.
- **Users** can manually define a collection of **Workspaces** in order to calculate a respective **CQ Annotation Score**.
- **Users** can export the underlying data used to calculate the **CQ Annotation Score** (i.e., the set of all relevant **Claim Instances** and **Question Instances**) for use in third-party curation or scoring services.

- While intended for representing an individual **User's** contributions, it can also be calculated using a collection of **Users** (where score would be a sum of the individual scores of the listed **Users**), or using a given **Workspace** (where score would be the count of claims which **Users** contributed as a member of the given **Workspace**).

Mapping Impact Score

The Mapping Impact Score is a metric intended to reflect the extent of a particular **User's** impacts on the network beyond their own contributions to identifying **Claim Instances** and **Question Instances**, such as their contributions to linking objects within the TrustFinder environment (e.g., adding **References** or **Claim Combinators**) within the context of a **Workspace** or a collection of **Workspaces**. Relevant objects include (i) **Claim Combinators**, **Question Combinators**, and **References** contributed by the **User** where they were the initial contributor, and (ii) the **Claim Combinators**, **Question Combinators**, and **References** which were contributed within or imported to a given **Workspace** or collection of **Workspaces**. The Mapping Impact Score (MIS) is defined as the cardinality of the set of **Claim Combinators**, **Question Combinators**, and **References** (referred to here as *edge objects*) formed from the intersection of the set of *edge objects* contributed by a given **User** (C), with the union of the sets of *edge objects* associated with a given collection of **Workspaces** (W).

$$\mathrm{MIS}(\mathrm{C}\,\{\,\},\{\mathrm{W}_{\mathrm{i}}\{\,\},{}_{\dots}\}) = |\mathrm{C} \cap (\mathrm{W}_{\mathrm{i}} \cup \mathrm{W}_{\mathrm{i}+1} \cup \mathrm{W}_{\mathrm{i}+\dots})|$$

- Every **Workspace** has a dynamically calculated Mapping Impact Score for each **User** which has contributed relevant objects within, either as members or as a result of imports from other **Workspaces**. This collection of scores includes scores for **Users** who are not members of that **Workspace**.
- **Users** can manually define a collection of **Workspaces** in order to calculate a respective Mapping Impact Score.
- **Users** can export the underlying data used to calculate the Mapping Impact Score (i.e., the set of all relevant **Claim Combinators**, **Question Combinators**, and **References**) for use in third-party curation or scoring services.
- While intended for representing an individual **User's** contributions, it can also be calculated using a collection of **Users**

(where score would be a sum of the individual scores of the listed **Users**), or using a given **Workspace** (where score would be the count of claims which **Users** contributed as a member of the given **Workspace**).

Network Impact Score

The Network Impact Score is a metric intended to reflect the impact of a particular **User's** impact via the invitation of other contributors into the TrustFinder environment within the context of a **Workspace** or a collection of **Workspaces**. Relevant objects include (i) **Claim Combinators**, **Question Combinators**, **References**, **Claim Instances**, and **Question Instances** contributed by **Users**, where they were the first contributor, and where they are members of the subject **User's** (i.e., the subject of the score) ***invitation tree*** (e.g., where the **User** was invited by an invitee of an invitee of the subject **User**), up to a distance of 6 degrees; and (ii) **Claim Combinators**, **Question Combinators**, **References**, **Claim Instances**, and **Question Instances** which were contributed within or imported to a given **Workspace** or collection of **Workspaces**. The Network Impact Score (NIS) takes as inputs a set of **Workspaces** (W) of length M, with each element representing a set of **Claim Combinators**, **Question Combinators**, **References**, **Claim Instances**, and **Question Instances** (referred to here as *contributions*) and a set of 2-dimensional vectors (u) of n length with each element representing a **User** within the subject **User's** ***invitation tree***, each vector contains (i) a set of **Claim Combinators**, **Question Combinators**, **References**, **Claim Instances**, and **Question Instances** that the element's respective **User** contributed (u_{i_c}) and (ii) the degree of separation of the element's respective **User** in the subject **User's** *invitation tree* (u_{i_d}). The Network Impact Score (NIS) is defined by (i) finding the cardinality of the intersection of the set of contributions by each given **User** in the subject **User's** ***invitation tree*** (u_{i_c}), with the union of contributions within the given collection of **Workspaces** (W); (ii) weighting the resulting cardinality by a parameterized decay function which takes the given **User's** degree of separation (u_{i_d}) as an input; and (iii) summing the results for each **User**.

$$\mathrm{NIS}(\mathrm{u}\,\{[u_{i_c}, u_{i_d}], \ldots\}, \mathrm{W}\,\{\mathrm{W}_{\mathrm{a}}\{\,\}, \ldots\}) = \sum_{i=1}^{n} |u_{i_c} \cap (\mathrm{W}_{\mathrm{a}} \cup \mathrm{W}_{\mathrm{a}+1} \cup \mathrm{W}_{\mathrm{a}+\ldots})|\, \frac{e^{-\frac{1}{u_{i_d}-1}}}{2}$$

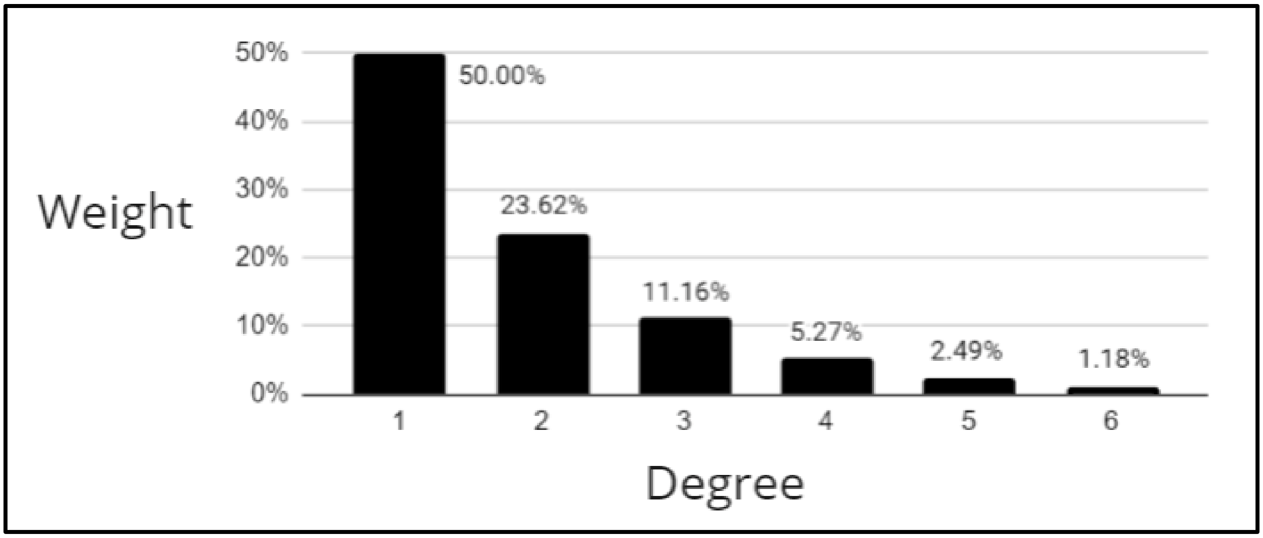


Figure 12. Degree of Separation impact on weight in Network Impact Score

Contribution Trust Score

The Contribution Trust Score is a metric intended, generally, to represent a particular **User's** or a **Workspace's** relative level of trust in another given **User's** annotation contributions (e.g., **Claim Instances**, **References**) or another **Workspace's** imported contributions. It reflects a **User's** or **Workspace's** expectations about the quality of annotation contributions by another **User** or **Workspace** (e.g., will this annotation contain errors?, will this individual use affordances as expected?). A Contribution Trust Score is set manually by a **User** for themselves or for their **Workspace,** and can be adjusted manually at any time. It is set on a scale between -1 and 1; where a rating of -1 is intended to represent a **User's** belief that the target of the rating would, without exception, purposefully or negligently contribute flawed annotations; and where a rating of 1 is intended to represent a **User's** belief that the target of the rating would, without exception, contribute properly formatted annotations, free of errors. This rating can be used to create filters on imports of annotation contributions within a **Workspace**, and will create visible indicators on presentation of annotations. **Users** and **Workspaces** must set a default rating to apply to unrated **Users** and **Workspaces**, upon registering or instantiation, respectively.

Given the subjective nature of the contents of annotations and the nature of expertise, **Users** can set conditional Contribution Trust Scores, which use a logical statements containing "and/or" combinations of **Entity Tags** and annotation types (e.g., IF (TAG$_X$ AND TAG$_Y$) OR Type $_{Reference}$) combined with a replacement rating. Where the logical statement holds true given the set of **Entity Tags** associated with a given annotation contributed by the target of the rating, the standard rating will be replaced by the defined replacement rating. This allows the marking of contextual trust, where, for example, a physicist's attempts to annotate **Claim Instances** within the domain of physics may be trusted at a higher level than their annotations related to psychology.

- Where there is more than one rating associated with the object, such as when there is both a personal rating and workspace rating, or where multiple conditional ratings triggered, the respective indicator related to the rating should be combined with others into new visualizations.
- **Users** should be encouraged to make conditional trust the norm via user experience mechanisms (e.g., by making conditional trust easy to assign via presented annotations with suggestions related to **Entity Tags** which are already present in the annotation).
- **Users** can export their Contribution Trust Scores for use in third-party curation or scoring services.

Assertion Trust Score

The Assertion Trust Score is intended to represent a particular **User's** or **Workspace's** relative level of trust in a **User's**, **Author's**, or **Publisher's** assertions. It reflects a **User's** expectations about the quality of **User Assertions** by a particular **User** or the quality of the contents of **Claim Instances** which are marked as *asserted* by the **Author** or **Publisher** of the **Content** in which they are found (e.g., does this person have a good grasp of the subject matter they are making assertions about? Is this person acting in good faith or are they being opportunistic?). An Assertion Trust Score is set manually by a **User** for themselves, or by a **User** for a **Workspace** and can be adjusted manually at any time. It is set on a scale between -1 and 1; where a rating of -1 is intended to represent a **User's** belief that the target of the rating would, without exception, purposefully or negligently assert false statements; and where a rating of 1 is intended to represent a **User's** belief that the target of the rating would, without exception, contribute objective and truthful statements. This rating can be used to create filters on imports of annotation contributions within a **Workspace**, and will create visible indicators on presentation of **User Assertions** and **Claim Instances**. **Users** and **Workspaces** must set a default rating to apply to unrated **Users**, **Authors**, and **Publishers**, upon registering or instantiation, respectively.

Given the subjective nature of assertions and the nature of expertise, **Users** can set conditional Assertion Trust Scores, which use logical statements containing "and/or" combinations of **Entity Tags** and annotation types (e.g., $\text{IF } (\text{TAG}_{\text{X}} \text{ AND } \text{TAG}_{\text{Y}}) \text{ OR Type}_{\text{Reference}}$) combined with a replacement rating. Where the logical statement holds true given the set of **Entity Tags** and annotation types associated with a given assertion, the standard rating will be replaced by the defined replacement rating. This allows the marking of contextual trust, where, for example, a physicist's assertions within the domain of physics may be trusted at a higher level than their assertions related to psychology.

- Where there is more than one rating associated with the object, such as when there is both a personal rating and workspace rating, or where multiple conditional ratings triggered, the respective indicator related to the rating should be combined with others into new visualizations and indicators.
- **Users** should be encouraged to make conditional trust the norm via user experience mechanisms (e.g., by making conditional trust easy to assign via presented annotations with suggestions related to **Entity Tags** which are already present in the annotation).
- **Users** can export their Assertion Trust Scores for use in third-party curation or scoring services.

Implications

The potential implications of affordances, social systems engineering mechanisms, and other aspects of the system are discussed below.

Local Governance

The solution space for managing governance, role and process, and mediation of conflict in human interactions online are extremely diverse, and best practices are highly dependent on local conditions. Any platform-level requirements and decisions reflected in complex or complicated definitions and rules for how users mediate conflicts, offer recourse, and manage roles and processes also create platform-wide threat surfaces with the potential for goal-blocking, inefficiency, and intrusions on community and user sovereignty. The recommended TrustFinder environment embraces this paradox as being reflective of reality, and makes the causative relationships explicit, opening up access to benefits from a more distributed and scalable approach wherein inter-community conflicts are managed via the formal structure of annotations that reveal the directional, and conditional, relationships between and among workspaces while intra-community conflicts remain in the purview of community self-governance. In this way, inter-community conflicts are effectively converted into community-oriented information differentials, the collective management of which yields value for all potential users. Specific platform-level governance affordances are recommended to be added only upon request by affected communities, and not required for use by all users across all communities. Given that workspaces can be arranged in complex import and export relationships by applying simple rules, many different, locally-adapted governance affordances may be facilitated without the need for specific standardized features (e.g., role-based access).

Empowering Communities and Users to Define and Assign Trust

Similar to the domain of governance, the solution space for managing reputation is extremely diverse, and user experience and quality control outcomes are subjective and highly dependent on local conditions. Any platform-level choice in complex or complicated definitions and rules for how user reputation is scored and impacted from behaviors of a user (or by the choices of other users) creates threat surfaces for misuse and counterproductive intrusions on individual and community level processes for deciding reputation. Further, it is not possible to create a curation or filter decision function that is free of bias, as curation and filtering is, by definition a discriminatory function. As such, any platform-level rules choice in defining curation and decision function for users will run a high likelihood of impacting users' trust in the system itself. As before, the discernment of this paradox reveals a system performance reality that is amenable to productive and value-creating management through community-based governance affordances, but in this latter case, directed toward individual reputation variables rather than the resource-focused attention of governance.

As discussed elsewhere, the goal of TrustFinder is to structure the information environment in order to enable users to find “trusted” sources of information. “Trust” is an emergent subjective internal state of a system (including “users” as a system), that is ultimately informed by elements that are external to the system. People and organizations that are empowered to discern (and measure) the degree to which performance of elements of a given system (or system component) are reliable and predictable may more confidently rely on the future performance of said system and come to “trust” said system in a mechanistic way. Users that have the capacity to identify and cultivate system elements that are relevant to their specific circumstances and upon which they can base such mechanistic “trust” have an advantage (in terms of cost and resource efficiencies) in leveraging and de-risking future interactions that is not available to others without such capacity. To this end, the users of TrustFinder specifically are empowered to define for themselves when and how to assign easily understood measures of trust (e.g., assertion trust scores, contribution trust scores), associated with other users, workspaces, and the authors and publishers of content, and to further specify in what contexts they apply and adjust those measures. It is recommended that TrustFinder take a facilitatory role in how researchers across disciplines adjust and access the values of these signals, as opposed to an authoritative role - and that it should be anticipated that its users will exercise agency to pursue their self-interest by self-binding to rules that offer reliability and integrity across a well-structured, navigable information system.

Rate Limiting Mechanisms on Spread of Trust

In complex information environments, trust may be counterproductively assigned using extrinsic signals such as affiliation and identity (or other surrogates for or abstractions of reliability and affinity) as opposed to intrinsic signals of quality and reasoning. While such assignment is understandable from the standpoint of interaction efficiency, when such a trust assignment is signaled publicly, the assignment will inevitably be affected by tribal dynamics and personal relationships and other agenda and contexts relevant to the users involved in later communications referencing such earlier trust assignments. In other words, the contextual foundations of the original abstraction of trust (e.g., to identity) is lost from the original communication, subjecting the naked communicated signal (data) to being interpreted by a later party in a different context (meaning) either through ignorance or malice, yielding so-called “mis” information and “dis” information respectively. TrustFinder makes it possible for researchers to manage communications to eliminate such “context stripping” of communications, by allowing them to manage trust signals privately.

The recommended TrustFinder environment benefits from an approach which stresses production of actor- and community-centric metrics (i.e., proximal, dynamic calculation from the perspective of a given user or workspace) which can be incorporated into more complex derivative down-stream curation and reputation analysis features by third-parties. The provision of services offering such down-stream insights have the potential to power new inter-disciplinary and trans-disciplinary insights in the academic sphere and new innovations in products, services, and markets in commercial contexts. The use of proximal calculation and presentation is applied as an alternative to a universal (i.e., platform-wide) or static reputation metric. This approach intends to limit the negative effects of context-stripped trust signals “going viral,” and to protect user and community ratings from being unduly affected by external pressures.

Scoping through Collaborative Work

Scoping the information environment through the use of mission-focused workspaces intended to facilitate collaborative work may affect the environment in a number of ways:

Subjectivity of Evaluation

Human knowledge is incredibly complex. In many cases (and contrary to what is often assumed) claims may only be “true” within certain contexts. For example, “home is where people will miss you when you are gone”, in some contexts, is a “true” statement, or a statement which “rings” true, or, at the very least, a statement which may be not helpfully marked as definitively false. It may not be the technical definition of a “home” from a given personal or cultural perspective, however it may be “literally” true in some cultural contexts, or “metaphorically” true within the context of a narrative analysis. This simple statement reveals the context dependency of the concept of “truth”.

By scoping the environment around collaborative work within a defined workspace, users can collaboratively refine their community’s information environment with the necessary context for user assertions, claims, and their relationships. Within the community workspace environment as contemplated here, members of the community do not need to ask for the permission or forgiveness of any outside party to apply a given set of context. They might be said to have “context/meaning sovereignty” within that information environment. Further, they can annotate and make assertions about claims applying their context-consistent elements without the need to fight for platform-wide consensus in order to enjoy the information environments that support them and enable them to perform work. While some may feel there is risk involved in allowing communities to define “their own truth”, a well designed system will be structured to make explicit the distinction of a contextual, community-bound “truth,” from a broader form of “truth” that is recognized across multiple contexts and multiple communities, which allows for the cultivation and management of dissenting views and innovation. In any event, fact-checking, censoring, or overriding the expressions of a given community that embraces a context bound, minority-position on a given “truth,” may be counterproductive. Generally, these kinds of interventions are only effective in terms of limiting effects of network exposure to undesired information or interpretation - but in the case of TrustFinder, said effects are already curtailed by the structure of workspaces.

Reduction of Information Overload

Any given text has the potential to include an overwhelming number of entities, claims, questions, and other annotations associated with it. The use of questions, claims, clusters of claims, and relationships between claims as a basis to scope workspaces improves the likelihood that the user will find annotations relevant to the task at hand.

Power Dynamics

Unbounded information collection activity results in cumulative build up of influence by committed contributors, and opportunities for “tyranny of the minority” phenomena, wherein small cliques get outsized control over what information in an environment is considered worthy of attention. With crowd-consensus mechanisms in place, the potential for tyranny of the minority is replaced by the potential for tyranny of the majority, where the interests of the majority truncate the interests of minority groups. The use of provisional and reconfigurable workspaces that can be selectively combined, abandoned, published, and republished by small teams allows for a freedom and flexibility that keeps both powerful cliques and homogenous crowds in check.

Neutral Discovery of Claims and Questions

Separating affordances for the discovery of claims from those that convey the opinions of users reduces the likelihood of tribal and affiliation-related dynamics and creates opportunities for common ground between groups with disparate interests and perspectives on the world. For example, two communities which vehemently disagree on the truth of a claim, can find common ground in the notion that “this article has an instance of this claim”; and even in cases of extreme disagreement, can at least agree on the title and citation metadata. This separation of concerns between different levels of analysis and complexity allows communities to benefit from each other's work despite their disagreements.

Modular and Flexible Construction of Claims Ecosystem

Traditionally, claims annotation is done on a document-by-document basis with a specific focus on the contribution of individual claims toward the argument a document is intended to advance. Allowing researchers to annotate the claims that are of value to their particular work simultaneously preserves quality of user-experience (i.e., not creating additional work for them unrelated to their current goals) and, as an incidental benefit of their self-interested annotation activities, also provides a modular, granular contribution to larger crowdsourcing solutions. As claims and references are linked to one another and are aggregated with the claims and references from other workspaces, small, individual contributions are brought together to create a rich, linked network of claims that no individual could have created alone. This is an example of familiar “network effects” of generating value, but here applied to meaning making across communities. Such emergent “meta-information” layers bear a relationship to baseline information similar to the relationship that meta-data has to baseline data, but in the case of such emergent, intercommunity context and meaning, situational awareness is extended to include formerly external components of context and meaning. Further, these relationships between claims can be represented as the key components of nearly any model of representation of argument and can be applied to any form of content (e.g., video, image, gif, text), which allows for advanced multimodal rhetorical analysis and reusability of claims information as training data in argument mining and artificial intelligence systems.

Claims as Networked Real-Estate: Gold Rush

Being the first to mark a claim provides both a first mover advantage on setting the tone and character for description and documents participation in its discovery. The reputational gains of being first, or more importantly, being first to provide a helpfully objective interpretation of a found claim, creates the opportunity for a “gold-rush” mechanism to drive adoption and participation. Further, given that reputation metrics are impacted by both the discovery and the annotation of claims, users are incentivized to perform high-quality claims discovery and annotation where it is most critical and valuable in both past and recent literature (e.g., finding and being associated with the discovery of claims which are at the root of a field are equally valuable to finding those which might be at the root of new fields or paradigm shifts). While such a mechanism can represent a risk to the intrinsic quality of annotation and encourage counterproductive rivalrous dynamics, there are several aspects of TrustFinder which are expected to keep these phenomena in check:

> ### Consumers of Found Claims are Incentivized to Merge
> The choice to merge two duplicate claims or to choose one annotated claim over another is now within the hands of those managing that workspace, and users are highly incentivized to detect and merge duplicate claims in the interest of reference stability. The incentive for rivalrous dynamics may increase with the value of the claim, but so do the incentives for maintenance of reference integrity.
>
> ### Competition
> Even where a user may intend to bury a rival's discovery in the interest of preserving their own status as the initial discoverer of a claim, and where they have control over a commonly referenced workspace, they do not have the affordances to maintain a control over the many other workspaces which may independently pull their rival's claims back in and merge them.
>
> ### Game Theory of Return on Work
> Given that reputation return for contributions is tied to the breadth of use and reference of the claim, in most cases, it will likely be a more reliable strategy to simply merge claims in order to increase likelihood of spread, even if it means a slight decrease in the perceived share of the reputation impact on use of a claim. The system rewards synthesis as much as it rewards discovery.

Use of Security Assurances

The affordances for annotation of personal opinions regarding claims found within content present threat surfaces for interpersonal aggression and intergroup tribal dynamics, and an opportunity for threat actors to use these vectors for purposes unrelated and contrary to the goals of the relevant community of users. As such, the TrustFinder environment requires users to engage with cyberphysical security measures in order to register an account in the system and to commit their assertions to the environment. This has several implications:

Cost of Engagement

The requirement to engage with security assurances in order to annotate assertions creates task-disrupting barriers that offer "shocks to consciousness" to the user to ensure they are unambiguously aware of the gravity of their interaction. This awareness is achieved via mechanism as opposed to being provided with disclaimers - users "experience" the weight of their decisions as opposed to simply being told about them and are prompted to consider the risk of their decision given the cost of engagement.

Cost of Entry

The use of security assurances creates a cost of entry to the environment that acts as hostile architecture to threat actors intending to make multiple accounts.

Separating Extrinsic from Intrinsic Rewards

Extrinsic rewards are those that have visibility from the outside (e.g., titles and status), and fungibility across people (e.g., material or currency), whereas intrinsic rewards are those that are inferred or experienced by a cognitive agent, such as personal fulfillment or a sense of purpose within a community. The potential for extrinsic and intrinsic rewards has significantly different impacts on behavior. Tendency to optimize toward extrinsic rewards is natural where they are offered, but this optimization axiomatically comes at the expense of the potential intrinsic value in the solution space. This being the case, creating simplistic extrinsic rewards for writing novels might generate *more* novels, though not necessarily better ones - and attaching "eyeball" or "dwell-time" related metrics, such as *how many people saw and liked my warning/endorsement*, will create perverse incentives for users to contribute what they believe the crowd will vote for, which may be in conflict with what they believe to be true.

The TrustFinder environment supplements its ability to support relatively modular, granular, narrow solution-space tasks (e.g., claims annotation) with extrinsic reputational rewards (i.e., CQ annotation score and mapping impact score, which reflect definable network impacts and use of contributions). Given the reliance on small-team focused workspaces, user assertions and responses to requests can be left to intrinsic reputational rewards - through the impacts users *feel that they make* on their local community.

Structure of Claims and User Assertions

The structure of claims paired with the attachment of user contributions to simple, self-reported levels of certainty and intensity enables the identification and application of new metrics about information integrity, opportunities for myriad forms of cognitive modeling regarding human engagement with clusters of claims and concepts, and opportunities to create related visualizations and accessible metrics for communicating status about integrity or informational conflict at the level of claim, document, or field (e.g., through the application of system status signals based on such things as color theory and simple summary statistics). Further, the highly structured relationships between claims and the structure of user assertions means that, where conflict arises, users are incentivized to engage in such conflict in a highly structured manner - resulting in hybrid information structures (i.e., composed of competing user assertions) which can be mined for insight regarding the volatility of certain claims. When using neutral claim annotations, as opposed to user assertions, users' interest in engaging in conflict (i.e., ensuring that claims they don't agree with are undermined, and that claims they agree with are supported) is harnessed as a driving force in mapping and connecting the rhetorical landscape as they search for supporting or refuting claims.

In addition, the flexibility of entity and custom tagging affordances in conjunction with open standards for interoperability with third party tools allows for communities to layer more advanced standards onto TrustFinder structures. For example, communities interested in more advanced rhetorical analysis of discourse are empowered to layer classification information onto objects, such as categories of claims (i.e., factual, definitional, causal, value, and policy) and other related data or categories of questions (e.g., interrogative, exploratory).

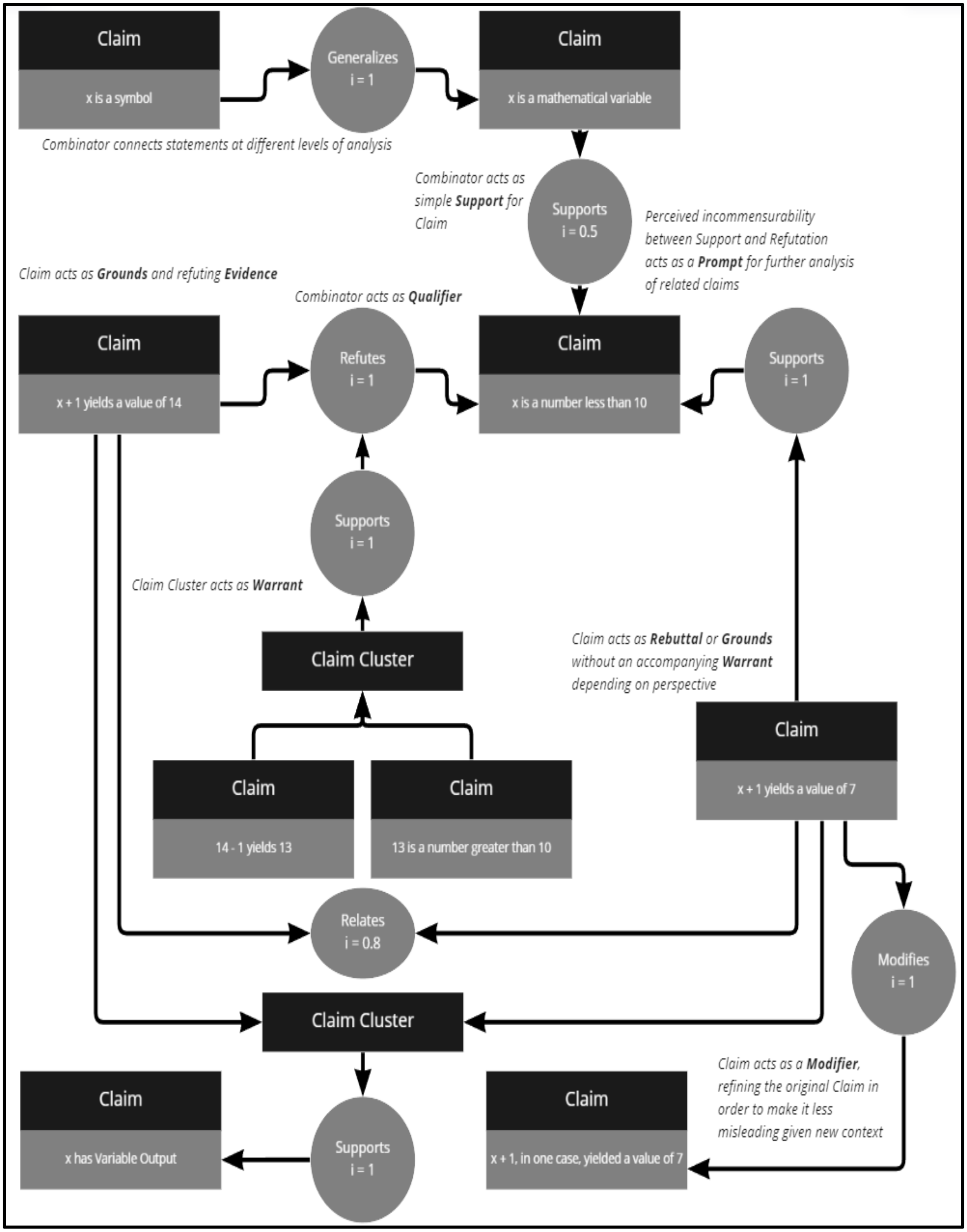

Claim
x is a symbol
Generalizes
i = 1
Claim
x is a mathematical variable
Combinator connects statements at different levels of analysis
Combinator acts as simple **Support** for Claim
Supports
i = 0.5
Perceived incommensurability between Support and Refutation acts as a **Prompt** for further analysis of related claims
Claim acts as **Grounds** and refuting **Evidence**
Combinator acts as **Qualifier**
Claim
x + 1 yields a value of 14
Refutes
i = 1
Claim
x is a number less than 10
Supports
i = 1
Supports
i = 1
Claim Cluster acts as **Warrant**
Claim Cluster
Claim acts as **Rebuttal** or **Grounds** without an accompanying **Warrant** depending on perspective
Claim
x + 1 yields a value of 7
Claim
14 - 1 yields 13
Claim
13 is a number greater than 10
Relates
i = 0.8
Modifies
i = 1
Claim Cluster
Claim acts as a **Modifier**, refining the original Claim in order to make it less misleading given new context
Claim
x has Variable Output
Supports
i = 1
Claim
x + 1, in one case, yielded a value of 7

Figure 13. Graphical representation of relationships between claims as a basis for representation of complex arguments, with example intensity ratings ("i") for claim combinators.

Compatibility with Other Systems

The structure of claims and references allows for import from (and the potential for export to) other systems which deal with claims discovery and reference management, such as Polyplexus, Swarmcheck, Paperpile, Mendeley, or Zotero.

Polyplexus

Polyplexus is a platform for crowdsourced collection of claims from documents and for hosting of claims-based exploratory research incubators. TrustFinder's claim instance and content objects would be highly compatible with Polyplexus' schemas, offering the potential for users to:

- import Polyplexus claims and driving questions in order to instantiate a workspace,
- export TrustFinder claim instances and reference data for upload to Polyplexus,
- export a TrustFinder workspace's claim instances, reference data, or claim clusters in order to submit claims to a Polyplexus incubator, or
- import claims associated with a Polyplexus incubator in order to instantiate a workspace.

Swarmcheck

Swarmcheck is a company which provides argument and discourse analysis and engagement tools for public and corporate use. TrustFinder's claim combinators and claim objects would be highly compatible with Swarmcheck's schemas, offering the potential for users to:

- import a Swarmcheck discourse map in order to instantiate a TrustFinder workspace, or
- export a TrustFinder workspace's claims and claim combinators in order to view and map discourse.

Reference Managers

Paperpile, Mendeley, and Zotero are platforms which provide reference management functions for researchers. TrustFinder's references and content objects would be highly compatible with most reference management schemas, offering the potential for users to:

- import and export reference objects.

Complex Knowledge Projects

The flexibility in creation of connections between and among workspaces allows for complex projects, constructed by multiple teams with separations of concern in workflow based on relevance of information. It also allows for individual researchers to find value even if they are isolated from all other users in the wider TrustFinder environment. Workspaces can be generated or populated with claims using queries of other workspaces to which they have access, and can have import and export integrations with other compatible systems, allowing for rapid synthesis in interdisciplinary, interorganizational work. Finally, TrustFinder workspaces can be used to help improve collection, accessibility, and dissemination of information resources for digital communities of practice at scale.

Gradients of Common Ground

Crowdsourcing solutions for information collection and interpretation can be difficult to implement when contributors don't share ontology or common narrative. The recommended TrustFinder environment assumes a wide diversity of viewpoints and implements a separation of concerns among objects to allow for communities which might disagree at one level of analysis to nonetheless cooperate on collection and analysis activities at another level where agreement is present (see Figure, "Gradients of Common Ground"). For example, two communities may have fundamental disagreements regarding the truth of a particular statement (i.e., at the level of user assertions), but can still agree on independent notions and issues such as the ideas and concepts involved and how they support or refute the statement (claim combinators), on where the statement is made (claim instances), and the relevant entities associated with the statement (stigmergic tags). In an extreme example, where two communities cannot even agree on the relevant entities associated with a given statement, they may, at the least, be able to agree on the name of a document or author (i.e., reference information). The use of workspaces with conditional import and export allows communities that would otherwise never interact to manage information sharing agreements that circumvent unnecessary conflict.

Mapping the Information Supply Chain

As of 2022, mapping the origin of a particular claim is a challenging, time-consuming task, even in literature with well-structured ontology and citation standards. While some reference mapping solutions exist, they are not necessarily accessible or sufficient for most use-cases, often contain errors, miss large swathes of relevant documents, and cannot keep up with the millions of new documents and datasets being generated each year. Further, even the best enterprise tools available rarely move beyond document-to-document links and references; it is only use-case specific tools, such as those found in legal study and practice, that offer affordances for semantic or conceptual provenance (e.g., precedent search). The recommended TrustFinder environment's reference and content objects, in conjunction with entity tags, claim instances, and question instances, allow for a collaborative mapping of implicit and explicit provenance of ideas across deep-time at the level of document and claims. Further, its flexible content object structure allows for claims of provenance to extend from the higher level of books all the way down to the more granular level of paragraphs, with attribution and reference annotation affordances that enrich and clarify context of citations and references appropriate for all such levels.

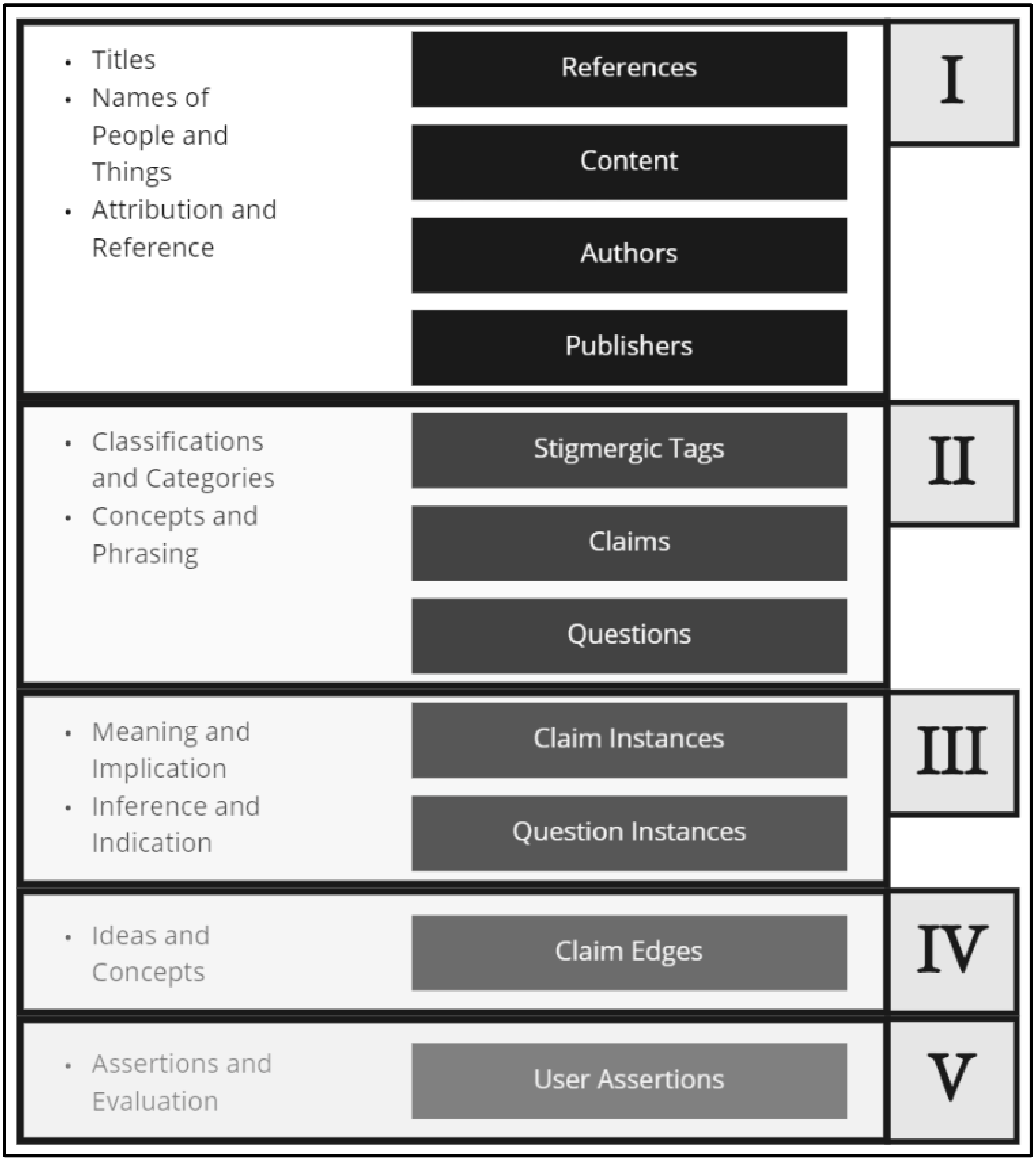


Figure 14. Gradient of Common Ground

EOS - Entity Oriented Search

The structure of the core TrustFinder objects, such as claims, claim instances, and content, allow for numerous queries that are driven by defined entities as opposed to syntax (i.e., language based search) which can illuminate implicit and latent relationships among claims and agents. For example:

By Content

A particular piece of defined content can be used as the object of search to yield:

- Claims within and their underlying claims.
- The content's implicit and explicit references.
- Other content which has a similar set of claims or references.
- Content which references the content used in search.

By Author

A particular author can be used as the object of search to yield:

- Common claims within their work.
- Common references they use.
- The claims they've made that aren't accompanied by their common refutations (e.g., what areas within their work might be biased or assumed).
- Publishers that have published their work.

By Publisher

A particular publisher can be used as the object of search to yield:

- Common claims within the work they publish.
- Authors they've published.
- How often they publish opposing points of view.

By Claim

A particular claim can be used as the object of search to yield:

- Content which presents or contains instantiations of that claim.
- Content which has instantiations of that claim primarily accompanied by refutations of that claim (e.g., to find critique articles).
- Content which has instantiations of that claim primarily accompanied by support of that claim (e.g., to find review articles).
- Claims which have certain relationships with the claim used in the search (e.g., supporting, refuting).

By Combinator Relationships

Combinator Relationships can be used as the object of search to yield:

- Search for claims within workspace that have very few combinator relationships to find potentially underexplored areas of research.
- Search for claims within workspace that have very high consistency in combinator relationships (e.g., claims with equal support and refutation) to find areas that may have been well researched but contentious.
- Exploration of the refinement of claims, by search and review of modification trees (wherein claims are refined through modification over time).
- Exploration of the generalization of claims, by search and review of generalization trees (wherein claims are generalized and specified across fields).

Infrastructure for Other Systems

The compatibility with external systems and the ability to create information "pipelines" between and among workspaces, in addition to enabling complex work, allows users to create ad hoc systems on top of TrustFinder.

Traditional and New Forms of Peer-Review

Journals and other research-publishing organizations could use workspaces to manage aspects of peer review that are concerned with claims and research questions, such as finding peer reviewers, evaluating the state of claims, and representing the rhetorical structure of the subject document. The ability to create multiple workspaces with conditional imports and exports means the potential for new forms of peer-review processes that are highly auditable and transparent, and allow for a larger number of participants.

OSINT SCADA

Organizations with high information collection and analysis requirements could use layers of interconnected workspaces to generate role-based information management and intelligence pipelines that can be contributed to at-scale and monitored in real-time. Given export and web annotation affordances, a collection of interconnected workspaces could be the basis for a supervisory control and data acquisition system (SCADA) for open source intelligence (OSINT) related purposes.

Technical Intelligence, Narrative Wargaming, and Exploratory Exercises

Users could build collections of interconnected, structured workspaces in order to engage in myriad narrative and technical intelligence related wargaming, collection, and exploratory exercises. For example, using separated blue (support), red (opposition), and green (communication) workspaces connected through intermediary workspaces with umpire-controlled selective disclosure. As another example, workspaces could be connected in order to allow for an adaptation of the “World Game” developed by Buckminster Fuller and others, wherein global resource availability and summary statistics are interactively and iteratively addressed by a collaborative team.

Background

Here, key frameworks and concepts are provided from works consulted and the works within this volume which guided the recommendations for the TrustFinder environment.

Argument Mining and Representation

Toulmin's Framework

The rhetorical framework of Stephen Toulmin has been used to make sense of and formalize argumentation and reasoning within myriad fields, including "science, law, management, art criticism, and ethics". The Toulmin rhetorical framework formalizes the structure of an argument through the relationships among 6 individual components:

Claim
The claim is the central *assertion* by an individual proposing an argument.

Grounds
Sometimes referred to as data, relevant facts, or evidence, the "grounds" of an argument is information that supports the claim.

Warrant
The warrant explains *why* the grounds support the claim. Warrants are claims themselves (often unstated assumptions) that must be accepted so that the original claim follows logically from the grounds. "Warrants confer different degrees of *force* on the conclusions they justify", which is communicated through a qualifier. A single argument (claim-grounds pairing) could be supported by multiple warrants.

Qualifier
The qualifier expresses the *relative strength* of the claim. It is often expressed rhetorically, through the phrases such as "might be", "probably", "certainly", or "axiomatically".

Backing

The "backing" component of an argument explains why the warrant has authority. The backing supports the warrant in the same way that the grounds support the claim.

Rebuttal

The "rebuttal" or counter-claim is a claim which refutes the claim or warrant.

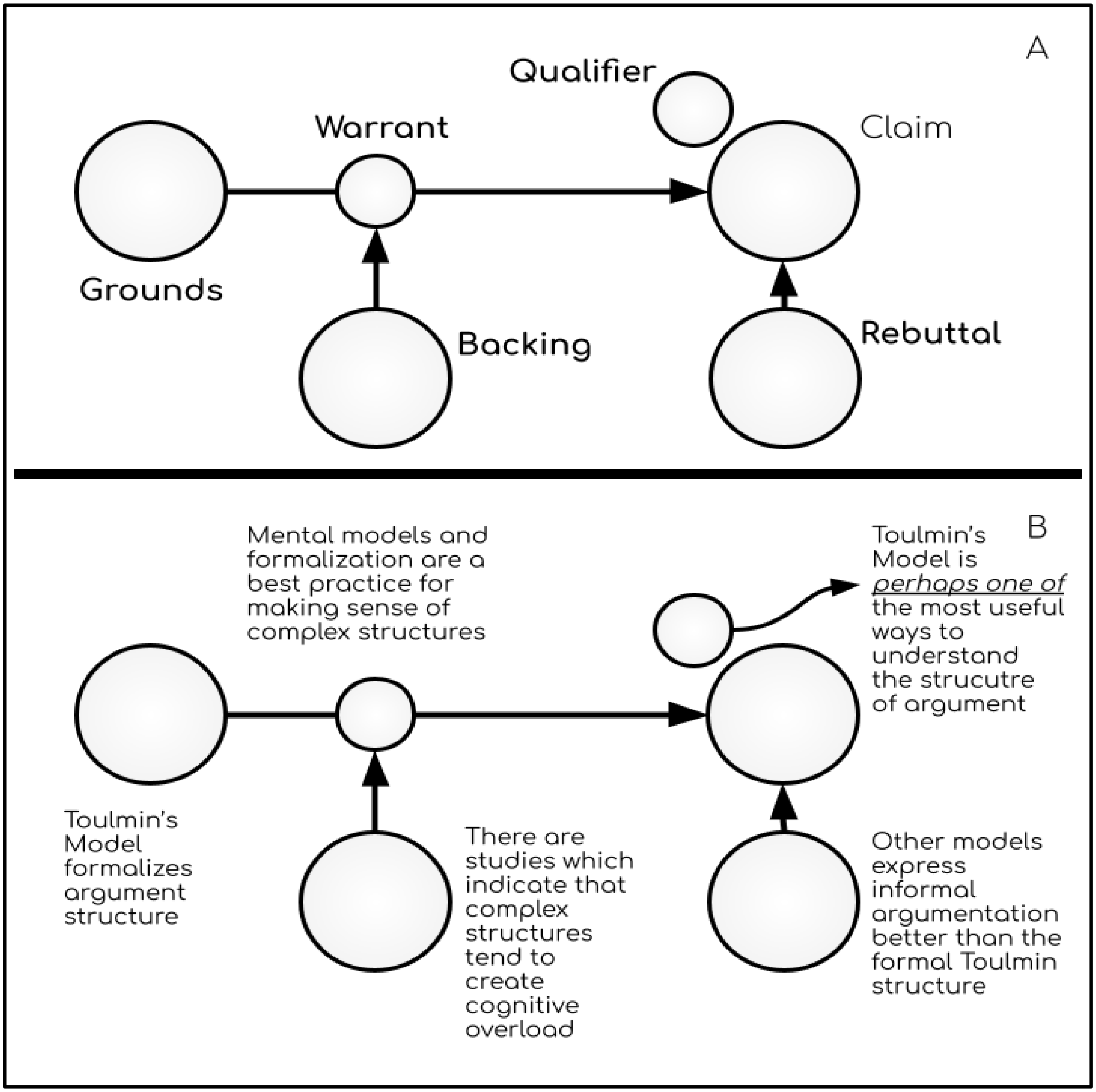

A
Qualifier
Warrant
Claim
Grounds
Backing
Rebuttal
B
Mental models and formalization are a best practice for making sense of complex structures
Toulmin's Model is *perhaps one of* the most useful ways to understand the strucutre of argument
Toulmin's Model formalizes argument structure
There are studies which indicate that complex structures tend to create cognitive overload
Other models express informal argumentation better than the formal Toulmin structure

Figure 15. (A) Toulmin's Model of Argumentation and (B) an example implementation

Toulmin himself asserted that this framework was not a “final” model for argumentation. Instead, it was the product of an exploration of the layout of argument driven by the intent to see *logic* developed into a formal science built on jurisprudence (legal philosophy). As such, it carries limitations, and has served as a foundation for myriad analyses and models which seek to address or overcome these limitations. It could be argued that chief among these limitations is addressing the interconnectedness of claims and their components - as the grounds, backing, and rebuttal attached to a claim can each be claims in their own right, and as such, have their own connected structures.

Stab and Gurevych Model for Argument Annotation

The Stab and Gurevych model for the annotation of argument is designed for extraction of granular and modular components of argumentation in persuasive essays. It is designed specifically for managing the relationships among claims and their support, refutations (attacks), and their own support or refutation for other claims. Of value here, is that this model uses a very simple set of rules and components in order to represent complicated arguments.

Statement

A statement is a piece of text which might contain components of argument and can be used as the basis for annotation.

Major Claim

The major claim is at the “center” of discourse, usually expressed rhetorically in the introduction of a piece of writing - indicating the author's stance on a particular topic.

Claim (Support or Attack)

This object expresses itself as *grounds* or *rebuttal* to the major claim by merit of the assigned “support” or “attack” relationship referred to as its “stance attribute”. A claim, like the major claim, is considered to be a “controversial statement” which will be supported or attacked within a text.

Premise

The premise supports (or attacks) the validity of a claim or major claim, or another premise by giving a reason.

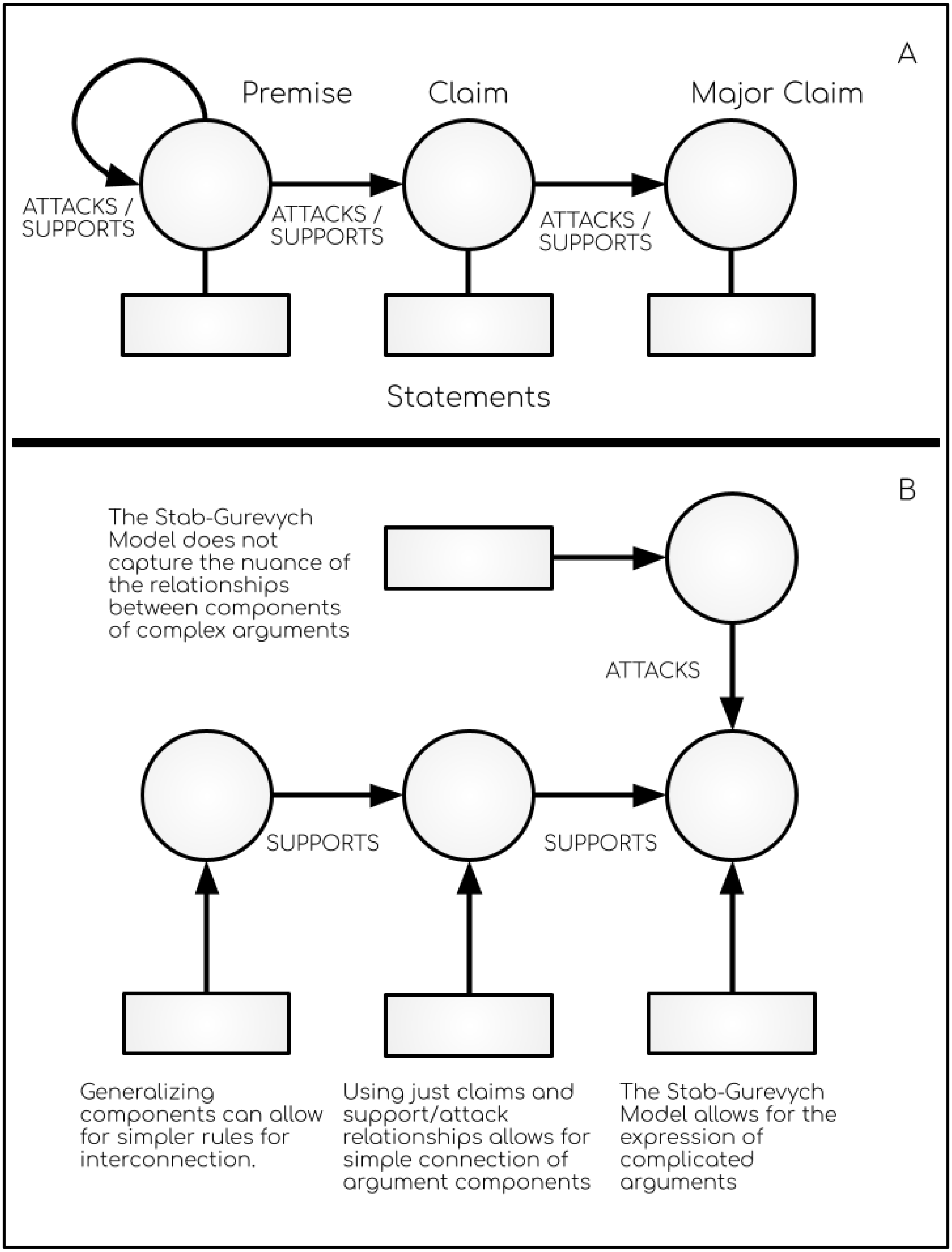

A
Premise
Claim
Major Claim
ATTACKS / SUPPORTS
ATTACKS / SUPPORTS
ATTACKS / SUPPORTS
Statements
B
The Stab-Gurevych Model does not capture the nuance of the relationships between components of complex arguments
ATTACKS
SUPPORTS
SUPPORTS
Generalizing components can allow for simpler rules for interconnection.
Using just claims and support/attack relationships allows for simple connection of argument components
The Stab-Gurevych Model allows for the expression of complicated arguments

Figure 16. (A) Stab-Guryvych Model for argument annotation and (B) an example implementation.

While this more general framework allows for complex interconnections between claims and helps analyze structured discourse, it, like the Toulmin model, comes with limitations. Some of these limitations could be interpreted to be a product of an intentionally constrained scope, as the work was intended only to advance the annotation of argument structures in a particular medium. For example, it provides no equivalent component to Toulmin's qualifier, and components cannot form relationships with the relationships between components (such as the warrant in Toulmin's model, which addresses the relationship between the grounds and the claim). Further, by merit of its focus on a major claim, it is best suited for annotating documents which are built via constrained writing tasks where all other claims sit in some hierarchy beneath the central claim.

Digital Rhetorical Ecosystem 3-Layer Model (DRE3)

The Digital Rhetorical Ecosystem 3-Layer Model or DRE3 model was designed to integrate rhetorical analysis with ecological theory in such a way as to make it compatible with a crowdsourced and computational analytics pipeline intended to produce a wide range of information products, such as publications and briefs, estimative and predictive metrics, and training data for automated analysis systems. It moves beyond rhetorical structure to consider object references and other content, and most importantly, is intended for analysis of argumentation communicated through multimodal content, with a specific emphasis on image memes. The DRE3 model does not structure argumentation so much as it structures the process of extraction of components and references within arguments embedded in content The purpose of this focus is to enable analysis of argumentation at the level of public discourse, or of argumentation within the context of a rhetorical ecosystem. The process of integrating an artifact (i.e., an image-meme) is expressed in 3 stages:

Entity Identification

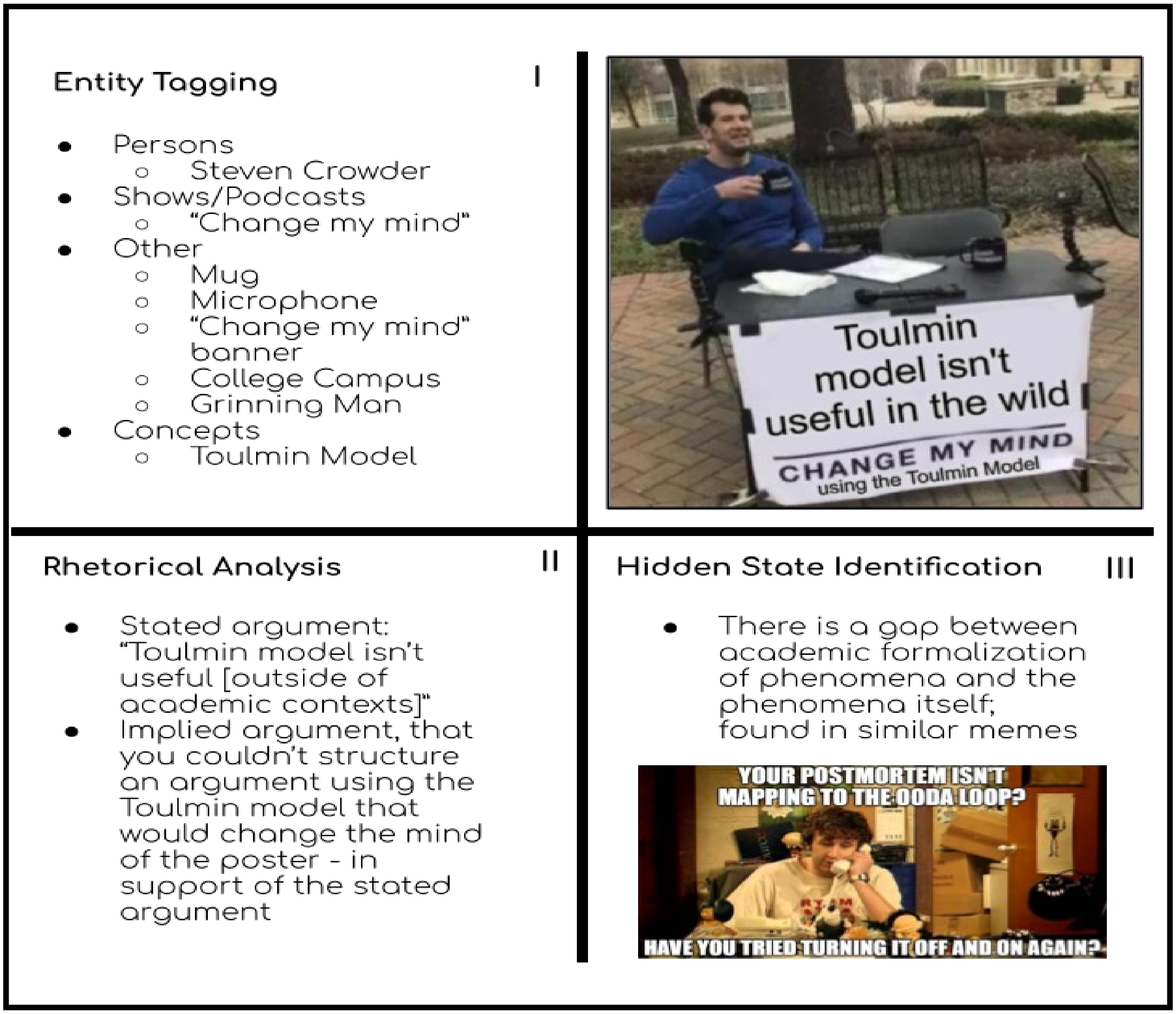
The first phase of DRE3 analysis is entity identification. In this phase, an analyst tags visible or implied entities, such as persons, organizations, locations, or concepts - enabling rapid collation of content with similar subjects. Further, it informs analysis in succeeding stages.

Rhetorical Analysis

The second phase of DRE3 analysis is rhetorical analysis. In this phase, an analyst decodes the relationships between the entities and their placement within the content. The objective is synthesis of these relationships into a central claim (or set of claims) made within the content.

Hidden State Identification

The final stage of DRE3 analysis is hidden state identification. In this phase, the analyst attempts to identify underlying broad claims which are implied by the claims within the content and by similar claims across other content.



Entity Tagging I

- Persons
  - Steven Crowder
- Shows/Podcasts
  - "Change my mind"
- Other
  - Mug
  - Microphone
  - "Change my mind" banner
  - College Campus
  - Grinning Man
- Concepts
  - Toulmin Model

Rhetorical Analysis II

- Stated argument: "Toulmin model isn't useful [outside of academic contexts]"
- Implied argument, that you couldn't structure an argument using the Toulmin model that would change the mind of the poster - in support of the stated argument

Hidden State Identification III

- There is a gap between academic formalization of phenomena and the phenomena itself; found in similar memes

YOUR POSTMORTEM ISN'T MAPPING TO THE OODA LOOP?
HAVE YOU TRIED TURNING IT OFF AND ON AGAIN?

Figure 17. Example implementation of DRE3 model

The DRE3 model, like other argumentation and argumentation analysis models, comes with its own limitations. For example, the extraction of hidden states and arguments is heavily influenced by the analyst, given the often esoteric and ambiguous nature of multimodal content. Its largest limitation may be that its value depends on the successful implementation of crowdsourcing solutions to annotate content, tag entities. and provide feedback on analyses.

Systems Design

Key Design Elements of Crowdsourcing Solutions

Attempts to solve problems, raise funds, collect evidence, or analyze data using large numbers of individuals is referred to as "crowdsourcing". Crowdsourcing solutions are deployed where automated approaches may not be effective or possible, and have been successfully deployed in a myriad of use-cases even where the crowd would not necessarily be perceived as competent in addressing the relevant solution space, such as using gamers to assist in the analysis of genetic and astronomical data. In this vein, crowdsourcing solutions have to be tailored to their use case, solution space, and crowd, resulting in a number of use-case specific categories of patterns of crowdsourcing solutions, such as prediction markets, where crowds are being used to predict events; or serious games, where games or game-like mechanisms are used in order to incentivize engagement or allow for a crowd to contribute to solution spaces for which they do not have the relevant competencies. Crowdsourcing solutions have to be carefully tailored to the conditions of their implementation for functional reasons, but also because of their dependence on engagement, it is difficult to make any single approach reliable - often, attaining reliability remains difficult even within a particular domain or use-case. Analyses of crowdsourcing solutions across the spectrum of use-cases suggest there are at least a dozen interconnected elements in common which contribute to likelihood of success, below these elements are compressed into three principles relevant to our purposes:

Task Communication

The system and users should have affordances to delineate, transmit, or broadcast task-related requests to others that are appropriate given the size of the crowd, diversity of the competencies of the crowd, complexity of the solution space, and number of requests that may be active at any given time. Difficulties in communications cost effort, time, and resources, and most importantly, impact both the likelihood of users attempting to solve tasks or their ability to broadcast tasks they cannot solve to others who can.

Task Solution Space

The solution space of tasks should have a complexity which is appropriate given both the competence and size of the crowd. The more agents involved in a solution space, the more modular, granular, specific, and well-defined the tasks and the measurement of their success must be in order for them to coordinate coherently. As an illustrative example, 100 people can come together to build a brick wall, but they cannot write a coherent novel. The more subjective the solution space is, and the less modular completed tasks are from one another (e.g., where each task impacts the solution space of the next), then the more individuals that are added, the more disagreement that will form within the crowd - contributing to incoherent results or lack of engagement. Where subjectivity in solution space is impossible to avoid, contributions must be well structured and as granular as is practicable.

Task Motivation and Feedback

The crowd should be given clear, relevant feedback about their interactions, and should have incentives which are appropriate given their competencies, the costs of performing tasks, and the potential impacts of incentives on outcomes. What constitutes relevant feedback or an appropriate incentive may, arguably, be more an art than a science - as some crowds may be effectively motivated and stimulated by feedback regarding their contributions to a community, whereas others may need more explicit incentives. However, incentives have to be tailored not only to the community but to the solution space itself, as extrinsic motivations such as currency or “points” can come at the expense of intrinsic motivation and therefore at the expense of the intrinsic value of the solution space. As an illustrative example, offering currency as a reward for producing 1000 words on a topic may be effective for generating words, but ineffective at generating value within them. Continuing with this example in order to illustrate the lack of standardized approaches across implementations: if individuals might have already been producing these 1000 words, and the currency was just a motivation for them to bring what they were already producing to the system, there is less risk of meaningless submissions, though moderation, reputation, and identity verification systems would still have to be put in place in order to reduce impacts on submission quality.

Coonradt’s Principles of Engagement

Coonradt, the “grandfather of gamification” asserted that activities which require extensive effort have 6 elements that must be present in order to be persistently engaging:

Clear Goals

The objectives of the work are clear and well scoped, making navigation toward those goals manageable.

Scorekeeping

The measurement of performance outcomes is clear, comparable, and unambiguous.

Feedback

Given the clarity of objectives and performance outcomes, individuals participating in a game or gamified system have reasonable basis to consider the impact of certain behaviors on results.

Choice

Games and game mechanisms provide players with choices, some clearer than others - the clearer the choices, the more valuable feedback becomes, and the more opportunities are provided for players to invest in understanding the impacts of their choices on outcomes and in innovating or adapting those choices.

Field of Play

The time and space in which the game is played are well scoped, so players have clear expectations entering this scope: they know what to expect, what is expected of them, and that the game will eventually end, and therefore that they will have time to rest if they exert themselves.

Skin in the Game

This concept from game theory was communicated to a much wider audience in the book of the same name by Nassim Taleb—that players need to acknowledge some value on the table, some potential cost or gain at stake that is tied to their performance in order to play effectively and fairly.

Key Principles for Social Systems Engineering

Social Systems Engineering (SSE) is concerned with the design of systems which involve or are driven by interactions between social agents. In traditional engineering, final system states can often be defined completely and provide highly reliable behavior through the use of (i) separations of concern among components, (ii) clear causal relationships and formal interfaces resulting in mathematically or algorithmically predictable phenomena, (iii) high reliability controls on interfaces, and (iv) predictably adaptive components with highly accurate feedback mechanisms. Humans have hidden states, hidden interests, and highly adaptive policy. As such, any system which includes human inputs will have a reliability which holds a nonlinear relationship with the degrees of freedom of said inputs and their impact on the system. Any system which has outputs that depend on the interactions between flexible human inputs is thus, by default, a complex system. The company AIE Nexus offers the following principles to help SSE clients define requirements and set expectations:

Simple Rules Create Complex Structures

Rules for interfaces and mechanisms should be as simple as possible, be moderated only by local conditions, result in modular and granular products, and rarely, if ever, contain exceptions. The relationships between the resulting granular products should be equally simple, and allow for flexible modularity in order to seed opportunities for the emergence of complex subsystems and structures.

You Cannot Design the Social System’s Mature State

For the majority of cases, you cannot predict from the starting state or from mechanisms or infrastructure what the resulting mature system will look like or if it will ever reach a mature state, even if a prior system had identical mechanisms and infrastructure and arguably equal starting state. While it is tempting to attempt rigorous definition and design of the mature state, the focus should instead be placed on requirements, controls, and standards which reduce likelihood of system failure and withdrawal of users, provide the users with value, control the structure of the systems outputs, and allow for iterative adaptation over time.

Retreatism and Withdrawal are the Default

Social systems implemented from scratch should have their mechanisms and rules designed with the assumption that new users are looking for a reason to leave until they have enough stake in the system to look for reasons to stay. Thus, the mechanisms and rules for interaction should be designed in such a way that individuals, by merit of use, are always accumulating stake in the system.

Harness Rebellion, Error, and Conflict

Assume that circumvention of the rules and use of the system’s human interfaces will be misused, abused, and rebelled against and that users will come into conflict. Do not assume that any component of the system is foolproof against any error or misuse. Instead, consider what adaptations or supplementary mechanisms can allow users or moderators to address or quarantine misuse and enable engineers to understand misuse in order to iteratively adapt the system over time.

Humans are Components in the System, Not Just Consumers

Social systems should be designed with the assumption that humans are "components" within that system, in addition to their roles as "users." With this expanded perspective, considering the "engineering" of human behavior (both as individuals and in their capacity as organizational actors) to increase reliability of outcomes becomes a default.

Meet the User Where They Are

Engineering user behavior or creating incentives from scratch is a perilous and generally unreliable process. Humans are not blank slates, and controlled environments with captured audiences can create misunderstandings about how game-theoretically-sound incentives may work in the wild. Wherever possible, mechanisms should be designed to harness, facilitate, and accommodate existing incentives, motivations, interests, processes, norms and expectations, and activities.

Trade-Offs are Inevitable, Prioritize Wisely

Every social system will be accompanied by trade-offs. For example, efficiency comes at the expense of reliability and quality and vice versa, and quality controls will negatively impact user experience in the short term in exchange for positive impacts in the long term. Trade-offs must be made explicit for participant evaluation, considered and prioritized carefully, and recognized as both unavoidable and amenable to co-management for enhanced system sustainability and resilience.

If Value to the User Depends on other Users, the System must be Seeded

If the value to the user depends on other users, then organic adoption in early stages is unlikely, as a lone user will likely not stay long enough to await the arrival of other users. A system must provide notable value to users in isolation or be seeded with inorganic users (e.g., paid users, stakeholders) in advance of achieving scale and maturity that is prerequisite to organic growth.

Clear, Meaningful Feedback is Good, Embodiment is Better

Clear, objective, and consistent feedback is standard practice for behavioral modification. However, wherever possible, behavior should be modified via affordances and structure to enhance reliability of system performance. For example, where users should exercise caution, it is more effective to implement affordances which require them to act out a process or ritual that requires caution or careful thought than it is to inform them to be cautious or to provide feedback where they failed to exercise caution.

MMOS Recipe for Serious Games

While there are numerous serious games designed for both research and education purposes, those implemented by the company Massively Multiplayer Online Science (MMOS) have been among the most impactful in the history of the field. To some extent, this success is due to their focus on finding ways to harness effort that is already being expended through existing activities, as opposed to building new activities entirely from scratch. The founders of MMOS have discussed a “recipe” for converting those individuals already engaged with digital activities into “virtually limitless computation engines for citizen science” [3]. An outline of this recipe, originally developed for use in the game EVE Online, is adapted for general use here:

Task Discovery

Find large-volume, modular tasks which require human annotation, analysis, or evaluation and cannot be effectively or reliably automated.

User Discovery

Find activities with which users with relevant competencies and capabilities are already highly engaged.

Task Mapping

Map the modular tasks to adaptations within the existing activities that harness or add to existing incentives while facilitating the performance of said tasks.

Theme Mapping

Make adaptations to the activity “aesthetically fitting and thematically adoptable” by the users.

Feedback

Make it clear to the users that by participating, they are making impacts beyond their own community.

Integration with Automation

Use the resulting data as training data for automated systems.

Active Inference Principles of Trust

The paper “Active Inference in Modeling Conflict: A Framework for Modeling Conflict in Business, Operations, Legal, Technical, and Social Contexts” presents 5 insights regarding trust and its impact on operations, informed by the Active Inference cognitive modeling framework. In conjunction with the ability to use ontology and formalization as a basis for behavioral engineering, these 5 insights can be argued to be principles for the design of collaborative systems:

Trust is Synonymous with Reliability

Trust can be characterized as a high level of certainty regarding the expectations of the policies and actions of another object, actor, or system. For example, we can trust a machine to function or not function, just as we could trust another person to act or not act.

Trust can be Externalized to Interfaces

Actors do not need to build trust with other actors if a higher level of trust can be assigned to an intermediary or interface through which they can instead engage. For example, we can externalize our trust to receive payment from a stranger to a payment system, as opposed to requiring trust in the stranger.

Trust can be Externalized to Symbols and Signals

Actors do not need to build trust with other actors if a higher level of trust can be assigned to symbols which reliably predict expectations about the environment. For example, “traffic signals allow drivers to externalize their trust to signals which inform the projection of other drivers’ behavior, as opposed to being left to develop trust with other drivers in order to share the road”.

Trust is a Prerequisite for Efficient Information Sharing

There are high costs associated with vetting information or sources of information, making communication without symbols, signals, interfaces, protocols, or pre-established personal trust cognitively expensive. Communication without externalization of trust or personal trust is axiomatically inefficient, either by merit of the costs of vetting, or the probabilistic risk of accepting low quality information or disinformation in lieu of vetting.

Trust is a Prerequisite for Collaborative Enterprise

In order to engage in collaborative enterprise, actors must have trust in relevant actors or externalize trust to a degree that is commensurate with associated risks.

Principles Related to Sustainability of a Commons

The study of "commons management" is rooted in the analysis and design of shared-resource systems, such as fisheries and grazing lands. While originally focused on natural resource management, commons management principles and research has found use in approaching other systems, with both real and abstract, or tangible and intangible resources, that encounter similar problems of common-resource use, such as conflicts over use, overuse, pollution, congestion, free-riding, unequal distribution, and availability of recourse. Hess and Ostrom, in their book, *Understanding Knowledge as a Commons*, provided eight principles for "robust, long-enduring, common-pool resource institutions":

Clear Boundaries

Where boundaries over what constitutes the common-pool being managed are blurred; responsibilities, needs, requirements, protocols, rules, and jurisdictional authority are blurred as well.

Rules are Well Matched to Local Needs

Empirical studies on common-pool resource governance have consistently indicated that "no single set of specific rules… had a clear association with success". Instead, rules needed to be adapted and adjusted to local requirements in order to sustain a resource commons.

Those Affected by Rules can Participate in Modifying Them

A commons "is often most efficient and durable when individuals affected by a resource regime" can participate in modifying its rules. This is in part because those who are affected are in the best position to understand how rules need to be adapted to map well to local needs, and more importantly are in the best position to understand what rules will be maladaptive or dysfunctional. Adaptive, sustainable governance systems tend to have the following characteristics:

- Information availability
- Recourse capabilities
- Rule compliance capability
- Rule-related infrastructure
- Preparation for and expectation of change

All of these characteristics require that rules be functional and well-mapped to the local environment and that those who are within the system participate in modifying them over time.

Right to Establish Local Rules

In order to enable rules which are mapped to local needs and avoid rules which generate dysfunction or encourage subversion, those affected by rules must be able to participate in modifying them. Those affected by rules cannot participate in modifying them if external authorities do not recognize their right to engage in establishing and modifying local rules.

Community is Empowered to Self-Monitor

Sustainability requires ongoing monitoring and evaluation. Those that are engaging in the interactions within the commons are in the best position to spot wrong-doing, negligence, or failure to meet standards.

Graduated System of Sanctions of Bad Behavior

Effective governance requires that there are “reasonable standards for small variations that [will] always occur due to errors, forgetfulness, and urgent problems”, and a graduated system of sanctions which become more severe to those “who do not learn’ from initial, more lenient encounters. The system itself also needs to graduate over time, increasing its severity and specificity. A governance system will often need to begin somewhat informally, as too many requirements for compliance too early can create disincentives for participation, and then develop over time into having more strict and clear sanctions for undesired behavior.

Simple and Low-Cost Mechanisms for Conflict Resolution

Conflict can provide opportunities for information discovery and refinement if facilitated and tempered in a controlled environment, in much the same way an engine produces work from heat. The goal of the governance system is not necessarily to end all potential for conflict, but to harness it to help the system as a whole reduce externalities and the potential for conflict to be destructive. Conflict resolution affordances need to be available, accessible, and affordable in order to avoid uncontrolled conflict.

Nested Enterprise

Sustainable commons tend to be those which have “nested enterprises” or those which have conflict resolution, monitoring, sanctioning, and other governance activities nested within a larger structure with “multiple layers” of activity and organizational components.

Infinite Games for Infinite Teams

The white paper “Infinite Games for Infinite Teams” introduced a role-based “case management [system] for knowledge mapping”. This system is expressed as a game which acts as a crowdsourcing solution for mapping narrative, arguments, and concepts together. The game begins with a “workspace” which is initialized with a “seed-meme”, such as “the central argument of a paper” or a hypothesis being investigated. The game has two modes, explore and exploit. In explore mode, “all team members can see all information”. In exploit mode, players then take on a role as either a Red, Blue, or Green contributor, each attaching concepts, documents, and arguments to the seed-meme.

Blue Contributor

Blue contributors take a defensive stance in making connections to the seed-meme, considering questions such as:

- What have previous thinkers/movements/stories done to counter this meme?
- How might the meme or narrative be instantly and transparently debunked?

Red Contributor

Red contributors take a more aggressive approach to contribution, considering questions such as:

- What would be an effective approach to changing people's mind, not just informing them or "raising awareness"?
- What is the most direct and devastating attack on the ignorance surrounding this topic? \

Green Contributor

Blue and Red contributors focus on evidence and logic, whereas Green contributors focus on "evocation of emotion, anecdotes, and narrative." Green introduces "kairos in the system, that is an understanding, sense, and sequence to the memes in a space", considering questions such as:

- How can ideas be communicated to multiple audiences?
- How might the same messaging be effective across audiences & media formats?

The contributions, when taken together, map an emergent, stigmergic memetic landscape. Disparate concepts from multimodal digital media are linked, providing a unique form of situational awareness around a topic.

Narrative Information Management

The paper, Narrative Information Management asserts that fields and specializations which intend to design and implement systems, protocols, and procedures to manage, synthesize, curate, and search digital information generally need to account for the provision of the following features:

Managing Information Gaps

The ability to recognize gaps in the knowledge base in order to direct attention, or to recognize gaps in personal knowledge and address them using an existing knowledge base.

Facilitating Situational Awareness

The ability to stay apprised or be notified of changes and updates in the relevant environment despite pressures of information volume, complexity, and rate of change.

Providing Descriptive and Explanatory Information

The ability to “dig” into particular components and objects for summaries and background information.

Compression

The ability to compress complex information structures using visualization, structure, collation, curation, ontological, and interactive mechanisms.

Case Management and Providing Prescriptive Information.

The ability to follow particular chained events or objects and be provided with actionable procedure-related information, such as best practices or next steps.

Synthesizing Intelligence

The ability to synthesize information in the knowledge base in order to generate new information products.

Facilitating Communication

The ability for users of the knowledge base to coordinate in a structured and coherent manner even where roles or expertise are heterogeneous.

Handling of Errors and Inconsistencies.

The ability for users to be directed toward and remediate errors and inconsistencies.

Management of Trust Signals

The ability for users to send, receive, assign, parse, and isolate signals of trust related to evaluation of information and of the intents and competencies of actors.

Social Systems Engineering

The ability for the system to adjust and modify behavior of the users in a way which promotes the health and sustainability of the system.

Framework for Synthetic Intelligence Guilds

The paper “The Synthetic Intelligence Guid: A Social Technology for a Digital Bazaar”, in proposing the foundations for a sensemaking-oriented community of practices, offers the basis for a number of generalizable prerequisites for decentralized knowledge management systems:

Prevent Race-to-the-Bottom and Rivalrous Mechanics

Mechanism design should prevent, address, or offset the impacts of Hobbesian, multipolar, and Thucydidean traps, coordination failures, negative-sum game theoretic dynamics, free rider, principal-agent problems, and other related dynamics.

Prevent Centralized Capture

Mechanism and underlying structure design should prevent, circumvent, or deincentivize the centralized capture or clique-control of any particular aspect of the system.

Shared Situational Awareness, Decision-Making, and Dissemination

Mechanism and underlying structure design should allow for and facilitate shared situational awareness of the information environment, support decision making activity, and allow for directed dissemination.

Clearinghouses

The system should provide simple clearinghouses for setting of information-related contracts and exchange of information products and services in order to break up silos and allow the flow of critical information between specialized groups.

Direction of Attention toward Opportunities and Gaps in the Knowledge Base

Mechanism and underlying structure design should incentivize search for and direct attention to opportunities and gaps in the knowledge base (e.g., “low hanging fruit”).

Domain-Specific Agents and Teams as opposed to Large Central Bureaucracy

As opposed to central bureaucratic structure, autonomous agents and teams should be incentivized and empowered to address challenges within the information environment.

Standards for Crowdsourcing and Crowdsourced Standards

The system should have structure and standards allowing for contributions at scale, and allow for the implementation, development, and spread of locally developed standards.

Group Transferable, Network Maintained Reputation Systems

Communities should be empowered to develop and manage local reputation systems with opportunities for information sharing between groups.

Right to Bundle, Buy, and Broker

Communities should have affordances to bundle, buy, and broker information products.

Funding and Acknowledgements

R.J. Cordes is funded through the NSF Convergence Accelerator Trust and Authenticity in Communication Systems Program (NSF 21-572), under award ID #49100421C0036 and is supported in research efforts through a Nonresident Fellowship with the Atlantic Council on appointment to the GeoTech Center.

Scott David is funded by the NSF Convergence Accelerator Trust and Authenticity in Communication Systems Program (NSF 21-572), under award ID #49100421C0036.

Daniel A. Friedman is funded by the NSF program Postdoctoral Research Fellowships in Biology (NSF 20-077), under award ID #2010290.

Works Consulted

In additon to the works included in the 2022 Volume “Structuring the Information Commons: Open Standards and Cognitive Security”

1. Xia H, Østerlund C, McKernan B. TRACE: A stigmergic crowdsourcing platform for intelligence analysis. of the 52nd …. 2019. Available: https://scholarspace.manoa.hawaii.edu/handle/10125/59484

2. Suhartono D, Gema AP, Winton S, David T, Fanany MI, Arymurthy AM. Argument annotation and analysis using deep learning with attention mechanism in Bahasa Indonesia. Journal of Big Data. 2020;7: 1–18.

3. Karachiwalla R, Pinkow F. Understanding crowdsourcing projects: A review on the key design elements of a crowdsourcing initiative. Creat Innov Manag. 2021;30: 563–584.

4. Goodnight GT. Complex Cases and Legitimation Inference: Extending the Toulmin Model to Deliberative Argument in Controversy. In: Hitchcock D, Verheij B, editors. Arguing on the Toulmin Model: New Essays in Argument Analysis and Evaluation. Dordrecht: Springer Netherlands; 2006. pp. 39–48.

5. Hitchcock D, Verheij B, editors. Arguing on the Toulmin Model: New Essays in Argument Analysis and Evaluation. Springer, Dordrecht; 2006.

6. Weinstein M. A Metamathematical Extension of the Toulmin Agenda. In: Hitchcock D, Verheij B, editors. Arguing on the Toulmin Model: New Essays in Argument Analysis and Evaluation. Dordrecht: Springer Netherlands; 2006. pp. 49–69.

7. Bianchini M, Gori M, Scarselli F. Inside PageRank. ACM Trans Internet Technol. 2005;5: 92–128.

8. Stab C, Gurevych I. Annotating Argument Components and Relations in Persuasive Essays. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland: Dublin City University and Association for Computational Linguistics; 2014. pp. 1501–1510.

9. Coonradt CA. The Game of Work. Gibbs Smith; 2007.

10. Friedman D, Applegate-Swanson S, Choudhury A, Cordes RJ, El Damaty S, Guénin-Carlut A, et al. An Active Inference Ontology for Decentralized Science: from

Situated Sensemaking to the Epistemic Commons. 2022. doi:10.5281/zenodo.6320575

11. Vyatkin A, Metelkin I, Mikhailova A, Cordes RJ, Friedman DA. Active Inference & Behavior Engineering for Teams. Active Inference Lab; 2020 Sep. doi:10.5281/zenodo.4021163

12. Afuah A, Tucci CL. Crowdsourcing As a Solution to Distant Search. AMRO. 2012;37: 355–375.

13. Ayaburi EW, Lee J, Maasberg M. Understanding Crowdsourcing Contest Fitness Strategic Decision Factors and Performance: An Expectation-Confirmation Theory Perspective. Inf Syst Front. 2020;22: 1227–1240.

14. Tamari R, Friedman D, Fischer W, Hebert L, Shahaf D. From Users to (Sense)Makers: On the Pivotal Role of Stigmergic Social Annotation in the Quest for Collective Sensemaking. arXiv [cs.SI]. 2022. Available: https://dl.acm.org/doi/abs/10.1145/3511095.3536361

15. He J, van Ossenbruggen J, de Vries AP. Fish4label: accomplishing an expert task without expert knowledge. Proceedings of the 10th Conference on Open Research Areas in Information Retrieval. groups.inf.ed.ac.uk; 2013. pp. 211–212.

16. Cordes RJ, Friedman DA. Narrative Information Ecosystems: Conflict and Trust on the Endless Frontier. COGSEC; 2021.

17. Friedman DA, Cordes RJ, editors. The Great Preset: Remote Teams and Operational Art. COGSEC; 2020.

18. Baxter G, Sommerville I. Socio-technical systems: From design methods to systems engineering. Interact Comput. 2011;23: 4–17.

19. Zadeh LA. Fuzzy sets. Information and Control. 1965;8: 338–353.

20. Maiers J, Sherif YS. Applications of fuzzy set theory. IEEE Trans Syst Man Cybern. 1985;SMC-15: 175–189.

21. McMillan J. Reinventing the Bazaar: A Natural History of Markets. W. W. Norton & Company; 2003.

22. Milgrom PR, Roberts J. Economics, Organization, and Management. Prentice-Hall; 1992.

23. Milgrom PR, North DC, Weingast* BR. The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. Econ Polit. 1990;2: 1–23.

24. Milgrom P. Auctions and Bidding: A Primer. J Econ Perspect. 1989;3: 3–22.

25. Glowacki A. From narratives of violence to narratives of peace: The renunciation of violence as a discursive phenomenon. 2013. Available: https://search.proquest.com/openview/12b2e7d7d9476024a27c66140ecfff86/1?pq-origsite=gscholar&cbl=18750

26. Cordes RJ, David S, Maan A, Ruiz A, Sapp E, Scannell P, et al. The Narrative Campaign Field Guide - First Edition. 1st ed. Cordes RJ, editor. Narrative Strategies Ink; 2021.

27. Sharma V, You I, Jayakody DNK, Atiquzzaman M. Cooperative trust relaying and privacy preservation via edge-crowdsourcing in social Internet of Things. Future Gener Comput Syst. 2019;92: 758–776.

28. citizensciencegames.com. List of Citizen Science Games. In: Citizen Science Games [Internet]. 2018 [cited 17 Jun 2022]. Available: https://citizensciencegames.com/games/

29. Dikopoulou A, Mihiotis A. The contribution of records management to good governance. The TQM Journal. 2012;24: 123–141.

30. Waldispühl J, Szantner A, Knight R, Caisse S, Pitchford R. Leveling up citizen science. Nat Biotechnol. 2020;38: 1124–1126.

31. Friedman DA, Cordes RJ. Infinite Games for Infinite Teams. The Great Preset: Remote Teams and Operational Art. COGSEC; 2020.

32. Cordes RJ. Games with serious impacts: The next generation of serious games - Atlantic Council. In: atlanticcouncil.org [Internet]. 21 May 2021 [cited 3 Jun 2021]. Available: https://www.atlanticcouncil.org/blogs/geotech-cues/games-with-serious-impacts-the-next-generation/

33. Friedman JA, Zeckhauser R. Assessing Uncertainty in Intelligence. Intell Natl Sec. 2012;27: 824–847.

34. Cordes RJ, Friedman DA. Emergent Teams for Complex Threats. The Great Preset: Remote Teams and Operational Art. COGSEC; 2020. pp. 1–15.

35. Waltz E. Quantitative Intelligence Analysis: Applied Analytic Models, Simulations, and Games. Illustrated edition. Rowman & Littlefield Publishers; 2014.

36. Waltz E. Knowledge Management in the Intelligence Enterprise. Artech House; 2003.