



AFRL-RI-RS-TR-2023-140

CAIRO-MS: CONFLICTING ACCOUNT INFORMATION RESOURCES IN OMNIVOROUS MEDIA STREAMS

LINGUISTIC DATA CONSORTIUM

JULY 2023

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

**AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (<http://www.dtic.mil>).

AFRL-RI-RS-TR-2023-140 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /

NANCY A. SEPULVEDA
Work Unit Manager

/ S /

MATTHEW J. KOCHAN
Technical Advisor,
Intelligence Systems Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

1. REPORT DATE JULY 2023		2. REPORT TYPE FINAL TECHNICAL REPORT		3. DATES COVERED	
				START DATE DECEMBER 2017	END DATE JANUARY 2023
4. TITLE AND SUBTITLE CAIRO-MS: CONFLICTING ACCOUNT INFORMATION RESOURCES IN OMNIVOROUS MEDIA STREAMS					
5a. CONTRACT NUMBER FA8750-18-C-0013		5b. GRANT NUMBER N/A		5c. PROGRAM ELEMENT NUMBER 62303E	
5d. PROJECT NUMBER		5e. TASK NUMBER		5f. WORK UNIT NUMBER R2FC	
6. AUTHOR(S) Stephanie Stressel					
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) LINGUISTIC DATA CONSORTIUM 3600 Market St STE 810 Philadelphia PA 19104-2653				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory/RIEA 525 Brooks Road Rome NY 13441-4505			10. SPONSOR/MONITOR'S ACRONYM(S) AFRL/RI		11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-RI-RS-TR-2023-140
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT In this project, Linguistic Data Consortium aimed to provide a comprehensive response to the need for AIDA development and evaluation data in four languages, covering text, image and video media and a variety of genres. Lasting contributions to the research community from this effort include multilingual multimedia corpora focused on specific topics of interest; manual annotation of entities, relations, and events using large type ontologies; resources for evaluating topic-level hypotheses from data; manual annotation of claim frames, an entirely novel annotation task; and resources to support similarity-based comparison of system and human annotations of semantically complex content.					
15. SUBJECT TERMS Text data, image data, video media data, hypotheses, ontologies, data annotation					
16. SECURITY CLASSIFICATION OF:				17. LIMITATION OF ABSTRACT	
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	SAR		18. NUMBER OF PAGES 78
19a. NAME OF RESPONSIBLE PERSON NANCY A. SEPULVEDA				19b. PHONE NUMBER (Include area code) N/A	

Table of Contents

List of Figures and Tables	iii
Abstract	v
1 Summary.....	1
2 Introduction	2
2.1 Related Work.....	2
2.2 Objectives.....	3
2.3 Technical Innovations and Challenges	4
3 Methods, Assumptions, Procedures	6
3.1 Scenario and Topic Development	6
3.1.1 Scenario and Topic Selection.....	6
3.1.2 Topic Model and Query Development	7
3.2 Data Collection.....	8
3.3 Ontology/Tagset Development.....	11
3.4 Within-document Annotation	13
3.4.1 Basic Structure and Approach	14
3.4.2 Changes in Scope and Exhaustiveness	15
3.5 Cross-document Annotation	19
3.6 Hypotheses.....	21
3.6.1 Phase 1: Facet-level hypotheses and prevailing theories.....	21
3.6.2 Phase 2: Perspectives	24
3.6.3 Phase 3: Claim Frames	25
3.7 Quality Control	31
3.8 Assessment.....	31
3.8.1 TA1 Assessment.....	31
3.8.2 TA2 Assessment.....	32
3.8.3 TA3 Assessment.....	33
3.9 Support for Similarity-Based Scoring	38

4	Results & Discussion.....	40
4.1	Scenario and Topic Development.....	40
4.1.1	Scenario and Topic Selection.....	40
4.2	Data Collection.....	46
4.3	Ontology/Tagset Development.....	52
4.4	Within-document Annotation	53
4.5	Cross-document Annotation.....	54
4.6	Hypotheses.....	55
4.7	Assessment.....	56
4.8	Support for Similarity-Based Scoring	58
5	Conclusions.....	59
6	Recommendations.....	60
7	References.....	62
8	Appendix A: Corpora Produced.....	64
9	List of Acronyms and Abbreviations.....	67

List of Figures and Tables

Figure 1: Example Ontologized Query	8
Figure 2: AIDA Multimedia Collection Pipeline.....	9
Table 1: AIDA Tagset Expansion for Phase 1	11
Table 2: AIDA Tagset Expansion for Phase 2	12
Figure 3: Saliency-Constrained Image Annotation Example	18
Figure 4: Unconstrained Exhaustive Image Annotation Example.....	18
Table 3: Entity Base Types Assigned in KB Linking	20
Table 4: Examples of Entity Coreference Clusters.....	21
Figure 5: Query and Facet-Level Hypotheses	22
Figure 6: Example Prevailing Theory with Events and Relations	23
Figure 7: Example KE Matrix	24
Table 5: AIDA Claim Frame Template.....	26
Figure 8: Unique Claims with Different Claimer.....	27
Table 6: Example Completed Claim Frame Template	28
Figure 9: Populated Claim Template Example	30
Figure 10: Clustering Duplicate Claims	30
Table 7: Cross-Claim Relation Type Examples.....	31
Figure 11: Graph Assessment Example	32
Figure 12: TA2 Zero-Hop KB Node (left) and Namestring (right) Queries	33
Figure 13: Matching Hypotheses to Perspectives.....	34
Figure 14: Evaluating Unmatched Hypotheses.....	35
Figure 15: Detailed Human-Readable Hypothesis.....	35
Figure 16: Full Phase 3 Assessment Pipeline for TA3.....	36
Figure 17: Identity Coreference and Coarse-Grained Clustering for Claims.....	38
Figure 18: Similarity Judgment Task for Tagset Concepts	39
Table 8: Phase 1 Topics.....	41
Table 9: Phase 2 Topics.....	41
Table 10: Phase 3 Topics	42

Table 11: Phase 1 Document Counts by Language	46
Table 12: Phase 1 File Counts by Media Type.....	46
Table 13: Phase 1 Document Counts by Topic	47
Table 14: Phase 2 Document Counts by Language	47
Table 15: Phase 2 File Counts by Media Type	48
Table 16: Phase 2 Document Counts by Topic	48
Table 17: Phase 3 Document Counts by Topic	49
Table 18: Phase 3 File Counts by Media Type	49
Table 19: Phase 3 Document Counts by Topic	50
Table 20: Unique Tags Annotated in by Phase.....	52
Table 21: Annotated Documents and Mentions by Phase and Practice/Eval Partition.....	53
Table 22: Similarity-Based IAA Scores for Entities, Relations, and Events	54
Table 23: KB Links and NIL Coreference Decisions by Phase.....	54
Table 24: Annotated TA3 Documents by Language	55
Table 25: Annotation Counts for Claims, Entities, Relations, and Events in TA3	55
Table 26: Count of Cross-Claim Relation Annotations by Type	56
Table 27: Phase 1 System Responses Assessed	57
Table 28: Phase 2 System Responses Assessed	57
Table 29: Phase 3 System Responses Assessed	57
Figure 19: Sample Lenient, Strict, and Crowd Semantic Similarity Judgments	58
Table 30: AIDA Data Legacy	59
Figure 20: Holistic Event Alignment by Human Assessors	61

Abstract

In this project, Linguistic Data Consortium aimed to provide a comprehensive response to the need for AIDA development and evaluation data in four languages, covering text, image and video media and a variety of genres. Lasting contributions to the research community from this effort include multilingual multimedia corpora focused on specific topics of interest; manual annotation of entities, relations, and events using large type ontologies; resources for evaluating topic-level hypotheses from data; manual annotation of claim frames, an entirely novel annotation task; and resources to support similarity-based comparison of system and human annotations of semantically complex content.

Multilingual multimedia corpora: Multimedia data has gained importance to meet the need for robust technologies that can extract key knowledge elements from diverse information streams. Most existing corpora focus on one media type or include multiple media types but treat each media file as a self-contained document. This is especially true for Ukrainian (a relatively low resource language), but also for Russian, Spanish, and English. The AIDA corpus presents a new addition to the available resources by providing multilingual and multimedia data designed to support extraction of complex information in a holistic way. Additionally, the AIDA corpora were designed to contain multiple and sometimes opposing narratives about specific topics, in order to support the development and evaluation of technology capable of building multiple hypotheses to account for alternative interpretations of data imbued with informational conflict. Corpora with these qualities are of special interest in the drive to develop technologies that can help identify and combat misinformation.

Annotation using large type tagsets: Understanding the nuances of conflicting accounts of events required a finer grained set of type tags to be applied to entities, relations, and events to support the systems' ability to reason over the data to construct alternative hypotheses. Over the course of the program, we expanded from a set of fewer than 100 annotation tags to the use of a Wikidata snapshot with many thousands of possible type tags.

Topic-level hypotheses: Over the course of AIDA, LDC contributed to the production of structured representations of topic-level hypotheses (called prevailing theories and perspectives at different phases of the program). These representations formed a set of knowledge graphs, each containing a different coherent account of a target set of events, which could be used to assess the correctness, completeness and coherence of system hypotheses. The prevailing theories and perspectives, as well as the assessment results containing judgments of system hypotheses, constitute a lasting resource that can be used for future development of systems that can produce coherent "stories" about events in the world in a structured form.

Claim Frames: The selection of the COVID-19 pandemic as a topic focus for the final phase of AIDA brought with it the need to carefully track not only what the entities, relations, and events comprised the conflicting accounts of aspects of the pandemic, but also who was making each of the claims as well as where and how they were made. This required an entirely novel annotation task known as Claim Frames. While the Claim Frame task was designed specifically for the Phase 3 AIDA tasks, it can also make a lasting contribution to the research area of factivity more generally and could easily be applied to new topics and/or datasets.

Similarity-based evaluation of events: In the final phase of AIDA, considerable program resources were devoted to supporting the ability to automatically score system output based on its semantic similarity to a human reference as well as for comparing independent annotations by multiple human annotators.

The data produced by LDC for AIDA makes a significant and lasting contribution to the resources available for tackling NLP problems in the spaces of conflicting information, misinformation, complex multi-event hypotheses and claim tracking. The multilingual and multimedia corpora provide a unique resource that is rich in conflicting accounts of complex topics, and the annotations cover over 1100 documents and represent well over a million annotation decisions. Most importantly, over a dozen new corpora resulting from AIDA will be published in the LDC public catalog starting in 2023. This data will be available to the research community at large, amplifying the government's investment in language resources by attracting new attention to the research and technology problems AIDA was designed to address.

1 Summary

In this project, Linguistic Data Consortium aimed to provide a comprehensive response to the need for AIDA development and evaluation data in four languages, covering text, image and video media and a variety of genres. LDC (1) located and harvested multimedia source data in English, Russian, Ukrainian, and Spanish containing conflicting accounts of events relevant to the scenarios and topics of interest; (2) created within-document annotation including entity, relation, and event annotation; (3) created a reference Knowledge Base (KB) updated with entities of interest and performed linking annotation using the reference KB; (4) created reference hypotheses in several different forms as program and evaluation needs changed; (5) provided manual assessment of system output to support program evaluations; and (6) delivered datasets in accordance with AIDA requirements.

In the context of the AIDA program, the linguistic resources produced by LDC were instrumental in enabling improved system capabilities for identifying conflicting accounts of events in multimedia data and constructing hypotheses as structured representations of those conflicting accounts. LDC produced 55 corpora for use by the AIDA community in system development and by NIST in AIDA and TAC KBP evaluations. The full set of corpora produced and distributed under AIDA appears in Appendix A; most of the data contained in these corpora will be published in the LDC catalog making the data broadly available for research.

Highlights of our work include:

- Produced multimedia corpora including four languages, totaling over 20,000 root documents and 90,000 individual media elements, plus over 875,000 annotation and assessment decisions
- Developed a novel annotation task (Claim Frames), and coordinated an AIDA community annotation effort
- Developed approaches to annotation with a large tagset, expanding from a set of fewer than 100 annotation tags at the start of the program to the use of DWD with many thousands of possible tags

2 Introduction

The DARPA AIDA program focused on the collection and annotation of data in multiple languages and modalities to support the creation of alternative hypotheses about events that occurred within specifically chosen real-world scenarios. LDC's efforts as the data team on AIDA involved collection of data in English, Russian, Ukrainian, and Spanish (three languages per phase, with Spanish replacing Ukrainian after the first phase). The data needed to contain known points of informational conflict on the topics of interest and needed to contain entities, events, and relations related to those topics that could be annotated in text, video, and images. The annotation of those entities, relations, and events required the creation of new larger annotation tagsets and annotation approaches to allow annotators to create consistent data using such a large tagset. The concept of a hypothesis consisting of multiple connected events and relations and their arguments as a coherent account of events in the world was another major new challenge requiring multiple iterations with DARPA, the evaluation team, and the entire AIDA community to arrive at task definitions that would meet DARPA's goals. The sections below discuss LDC's objectives and the technical challenges involved in meeting them in more detail, followed by sections describing the methods, results, and conclusions of this work.

2.1 Related Work

The AIDA corpora make several contributions to linguistic resources in the languages they include, especially due to the multimedia nature of the corpus, and the focus on conflicting information.

Along with English, Russian (all phases), and Spanish (phases 2 and 3), AIDA also includes data for Ukrainian (phase 1), which is a relatively low resource language. Shvedova (2020), in introducing the General Regionally Annotated Corpus (GRAC) of Ukrainian (uacorpora.org), points out the lack of available resources for Ukrainian. The handful of published Ukrainian corpora in existence prior to AIDA are text-only resources (e.g., Tracey and Strassel, 2020; Sitchinava, 2012; Grabar and Hamon, 2019; Grabar et al., 2018).

Multimedia data -- especially documents on the web that contain embedded images, video and audio clips, infographics, social media snippets and reader comments interwoven with traditional text -- is gaining importance as sponsors seek robust technologies that can extract key knowledge elements from diverse information streams. Especially for Ukrainian, but for Russian, Spanish, and English as well, most corpora focus on one media type or include multiple media types but treat each media file as a self-contained document. Exceptions are mostly English corpora of subtitled video or images with captions (e.g., Han et al., 2018; Lin et al., 2015). The AIDA corpus presents a new addition to the available resources by providing multilingual and multimedia data designed to support extraction of complex information in a holistic way.

Finally, the subject of misinformation and conflicting narratives is of great interest and importance in the global information age (Gottlieb, 2020; Dan, 2021; Shu, 2020; Szostek, 2018). The AIDA corpora were designed to contain multiple and sometimes opposing narratives about specific topics within each scenario domain, in order to support the development and evaluation of technology capable of building multiple hypotheses to account for alternative interpretations of data imbued with informational conflict.

2.2 Objectives

The overall goal of the AIDA data effort was to provide data to support system development and evaluation for TA1, TA2 and TA3. As specified in the design of the program, and re-emphasized by DARPA on numerous occasions, the development data provided to performers in TA1, TA2, and TA3 was meant to provide examples that were representative of the data structures and annotation approaches, but not to provide large-scale training data sets.

Within this broader goal, LDC had several specific objectives.

Task 1: Project Management, IT, Distribution

Objective: To provide coordinated oversight of all AIDA TA5 tasks including centralized data/annotation pipelining; secure storage, backup, mirroring, monitoring and related infrastructure; secure intellectual property rights as required to permit broad dissemination of resources within and outside of AIDA; design annotation and collection requirements and QC methods to support integrated multimedia approach; share knowledge within the program and the broader research community as appropriate; satisfy all program reporting and related requirements

Task 2: Build Topics and Queries

Objective: To support collection and annotation of data relevant to the scenario for each phase; ensure adequate coverage of scenario topics; ensure presence of conflicting knowledge elements to support generation of multiple hypotheses

Task 3: Data Collection

Objective: Collect a corpus of at least 100,000 documents¹, comprising formal and informal data sources across a range of media types (text, speech, video, image) to provide suitable material for AIDA system evaluation.

Task 4: Label Within Doc Knowledge Elements

Objective: To provide labeling of knowledge elements including entities, events/sub-events/actions, relations, location, time, and sentiment to support evaluation of AIDA systems in TA1

¹ Note that during Phase 1, target data volumes were reduced by mutual decision of the AIDA research community, NIST and DARPA. Processing times required for performers to ingest the larger corpus exceeded time available during evaluation cycles and so a decision was made to incorporate richer annotations on somewhat less data. Final corpora for each phase range between 19K and 44K individual media and text files.

Task 5: Label Cross Doc Knowledge Representation

Objective: To provide linking of knowledge elements to a knowledge base to provide disambiguation and aggregation in order to support evaluation of AIDA systems in TA2

Task 6: Create Hypotheses

Objective: To provide human generated hypotheses to support evaluation of AIDA systems in TA3

Task 7: Assessment

Objective: To provide assessment of AIDA system output against ground truth

2.3 Technical Innovations and Challenges

Supporting the objectives listed above presented several challenges that required significant innovations. One of biggest technical innovations in the data realm for AIDA was the integration of multiple genres and modalities in a unified collection and annotation scheme. Past projects had largely maintained a separation between different genres, domains and media types. While recent programs like DEFT (DARPA, 2012) and LORELEI (DARPA, 2014) had required a unified treatment of disparate text data genres, defining a similarly cohesive approach for different media types was a novel challenge requiring innovation in nearly all parts of the data pipeline. Beyond new collection techniques, AIDA also required significant innovation in defining procedural approaches to annotation in order to ensure consistent decision-making across media types. For instance, while a named entity mention in text is often sufficient to uniquely identify an individual in the context of a specific scenario, it can be more difficult for annotators to establish unique identity for entities in other media. Detailed information about the approach to data collection can be found in section 3.2, and the results of collection are summarized in section 4.2)

An additional area of significant innovation for AIDA was the generation and assessment of hypotheses. The production of multiple human-generated accounts of events that pull together events and relations from different documents, across languages and media types, and are sufficient to serve as examples of the kinds of hypotheses systems should produce from the data was a new type of task and presented the added challenge of shifting in its definition and evaluation as the program evolved from phase to phase. In the first two phases of AIDA, system hypotheses were represented as a connected subgraph of events and relations and their arguments that were related to a specific topic within the scenario and were responsive to “Statement of Information Need” (SIN) provided as a prompt. The task of building a set of human-generated hypotheses for comparison involved considerable collaboration between LDC and MITRE, who served as subject matter experts for AIDA and were involved in the selection of scenarios and development of topics. Using MITRE’s domain knowledge and LDC’s knowledge of the data and annotation, various forms of hypotheses were created throughout the phases of the program, shifting either slightly or dramatically with each phase to accommodate new understandings of system capabilities and changes in evaluation design. Over the three phases, including the pilot effort, hypotheses took the form answers to queries addressing single facets of information (facet-level hypotheses in Phase 1), collections of events and relations required to instantiate a more complex multi-faceted query

answer along with a natural language representation of the hypothesis (prevailing theories in Phase 1 and perspectives in Phase 2), and finally claims about COVID-19, including natural language and structured representations of information required to instantiate the claim along with information about who made the claims and cross-claim relationships between claims. Details about the development of hypothesis information can be found in section 4.6, and a summary of the results in section 5.6.

The challenge of building these hypotheses for each new scenario and topic was related to another of the major technical challenges for the data team in AIDA: defining a set of event, relation, and entity types that would adequately cover the content of each new scenario with enough fine-grained detail to enable systems to produce hypotheses consisting of labeled events and relations that would distinguish the differing accounts of what happened. The need for a very fine-grained set of annotation labels to support the hypothesis task required considerable effort to define both the set of tags and the annotation procedures to enable use of such a large tagset compared to previous efforts. The expansion of the tagset from a small set of defined labels to a much larger set with careful definitions, constraints on allowable arguments, and customized templates for annotation in the Phase 1 between the pilot effort and the final Phase 1 evaluation represented a four-fold increase in the number of types available for annotation, and additional changes in Phase 2 expanded the number of argument roles for many events, and allowed events to serve as arguments of other events, marking a significant addition of complexity to the structure of the annotation. Finally, in Phase 3, the program's shift to using Wikidata increased the number of available types into the thousands and required additional innovations in both annotation (to incorporate the use of this more open ontology with less carefully curated structure) and assessment (to enable evaluation of semantically similar but potentially structurally different representations of the same events and relations as correct). Details about ontology expansion can be found in section 4.3, and a summary of the resulting annotation tags in section 5.3. A summary of work conducted to aid in the understanding and scoring of semantic similarity across differently labeled claims is discussed in section 4.9.

3 Methods, Assumptions, Procedures

Creation of data for AIDA was a multi-stage process. The first stage was development of scenario and topic models to guide all subsequent stages of data production, from data collection through assessment. The development of scenario and topic materials involved identifying potential information needs related to the scenario and compiling preliminary information about alternative accounts of events that were expected to occur in the data. This scenario and topic information was then used to inform collection of suitable data to support the scenario. In the next stage, a subset of collected documents relevant to the scenario are labeled for salient knowledge elements (events, relations, and entities). The annotated knowledge elements are then disambiguated across the corpus by linking to a knowledge base and/or performing cross-document co-reference. The final stage of the annotation pipeline was to produce and/or annotate hypotheses. The definition of what constitutes a hypothesis, and consequently the methods used to produce them, changed significantly from the start of the program to the final phase, but the core concept remained the production of a human-generated set of alternative accounts of scenario-relevant events that could be used to evaluate system performance. The final stage of data development for each phase of AIDA was the assessment of system output for the phase evaluations.

This section describes the methods, assumptions, and procedures for each component of the AIDA data development effort. Where there are significant differences across the three phases of the program, those are described in detail.

3.1 Scenario and Topic Development

3.1.1 Scenario and Topic Selection

For each phase of AIDA, MITRE and DARPA proposed one or more candidate scenarios of interest. LDC then did a preliminary survey of data availability to provide feedback, including information about availability of non-English language data and annotators before scenario selection was finalized. The final scenarios selected were as follows:

- Phase 1: Russia-Ukraine Relations
- Phase 2: Crisis in Venezuela
- Phase 3: COVID-19 Global Pandemic

Within each broad scenario, LDC and MITRE worked jointly to develop specific topics to serve as a focus of data collection, annotation, and evaluation. In all three phases, MITRE developed a scenario document that included detailed descriptions of potential topic areas and known informational conflict for each candidate topic. LDC and MITRE worked jointly to select and develop a list of final topics that included diverse informational conflict, with LDC providing information about criteria such as prevalence of differing assertions about what happened, availability of data in all three target languages, and availability of multimedia data. LDC also provided input about the key events and concepts involved in the conflicting accounts for each topic with regard to whether they were already covered by the existing set of types in the entity, relation and event tagset, and identified crucial concepts that could not be covered by the ontology tagset. The work involved in adding new types to the tagset is covered in section 4.3 below.

Final selection of topics was made jointly by LDC and MITRE to include those which had good data availability and multiple aspects of informational conflict. We also considered which kinds of informational conflict could be expressed readily in the annotation tagset, whether by existing types or by creating new types. As a final step, LDC and MITRE partitioned the topics into practice and evaluation sets, ensuring that tagset types that might be needed for eval topics also appeared in practice topics.

Information about the results of the topic development process, including specific topics selected, is provided in section 5.1.1.

3.1.2 Topic Model and Query Development

For each topic, LDC created a topic model to be used as a reference by annotators during data collection and annotation, to help guide their decisions about topic relevance. The topic model included the topic ID and title, a list of natural language queries about the topic, and two kinds of topic descriptions. The natural language queries were questions focusing on a single point of informational conflict (e.g., “Who carried out the drone attacks in Caracas on August 4, 2018?”). The first description was a summary of the key points of the topic from a neutral perspective. This description sought to contextualize the information with which annotators should be concerned and to which they might be exposed during data collection and annotation. The second description detailed the points of informational conflict intended to be the focus of the topic, as well as the known conflicting narratives related to each point of conflict. This was not intended to be an exhaustive account of the conflicting information related to a topic, and annotators were free to explore any conflict not mentioned in this description during collection and annotation. All descriptions were based on a combination of information from MITRE’s scenario document as well as documents gathered during preliminary data collection conducted during the topic development process.

The topic model documents served to orient the annotators to the topics and highlight specific points of informational conflict during data collection and annotation, prompting annotators to find documents that contain enough information to answer the queries and thus ultimately support the construction of hypotheses. In addition to supporting the annotators’ work, the queries were also used to support NIST in developing “Statements of Information Needs” (SINs) in the evaluation of systems. The queries were developed to fully cover the expected informational conflict in the topic. Most queries asked about core who/what/where/when aspects of the conflicting information in the topics, but some queries also addressed richer context, for example, by asking “why” questions.

The queries were developed as natural language queries, but an additional version of the queries was provided to NIST with additional event/relation (and argument) information added. NIST used the event/relation types and argument structures to construct machine-readable SINs for the evaluation that would address the points of informational conflict. We referred to these queries with added type and argument structure information as “ontologized queries”.

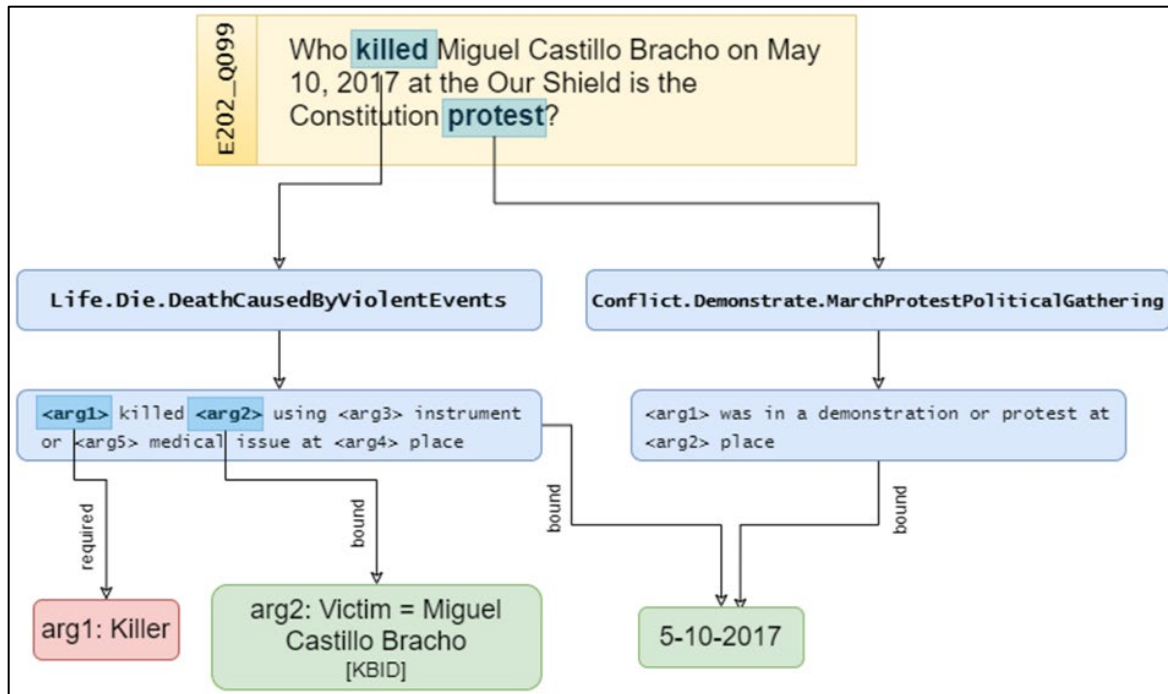


Figure 1: Example Ontologized Query

In order to create an ontologized query, annotation tagset types were assigned to the events in the natural language query. Argument roles of the events were added, along with the argument fillers from the natural language query, and each role was labeled as either required (the answer to the query that systems were required to provide from the data for evaluation purposes) or bound (information drawn from the natural language query itself). In the example shown in Figure 1 above, the required argument role is the killer, and the bound roles provide the information to the systems that the killer of interest is the one who killed a specific victim (Miguel Castillo Bracho) at a specific place and time (the protest event on May 10, 2017).

3.2 Data Collection

The source data for AIDA consisted of multimedia (text, image, and video) data in English, Russian, Ukrainian, and Spanish. English and Russian were collected for all three phases, while the second non-English language changed from Ukrainian to Spanish after Phase 1. A portion of the documents were directly relevant to the specific topics and queries discussed in section 3.1, while other documents were generally relevant to the broader scenario for each phase without addressing the particular topics of interest or were background documents of unspecified topic content.

To ensure that sufficient content to support annotation and evaluation was present in the corpus, we seeded the corpus with documents specifically collected for their relevance to the topics and queries. LDC annotators used the topic model to guide their search for documents on the web, with the goal of locating a diverse set of documents with respect to media types, languages, and conflicting perspectives. Annotators were instructed to search for documents that addressed the topics' queries across multiple different types of media and supported multiple competing or conflicting answers to the topics' queries. For each document they located, annotators collected

information about the URL, language, presence of relevant text, image, and/or video, and presence of query answers.

Beyond the manually scouted topic-relevant documents, additional data was collected automatically from websites that were rich in appropriate data to fill out background documents required for the corpus. Keyword ranking was used to ensure that there was sufficient scenario-relevant content on the automatically harvested data to support program requirements.

URLs identified in data scouting, as well as full site information for the automatic background collection, were fed into LDC's multimedia data collection pipeline, which collected all text, image, and video elements of the specified web page, processed them into separate files, and recorded metadata maintaining the association between the URL and each element of the page, as well as additional information such as publication date, download date, and relative position of each element on the original web page.

The LDC data collection pipeline is designed to ensure suitability of data for downstream annotation and evaluation, with the goal that data developed in AIDA should be broadly shareable outside the program (e.g., in open evaluations, publication of corpora).

Figure 2 illustrates the data collection pipeline.

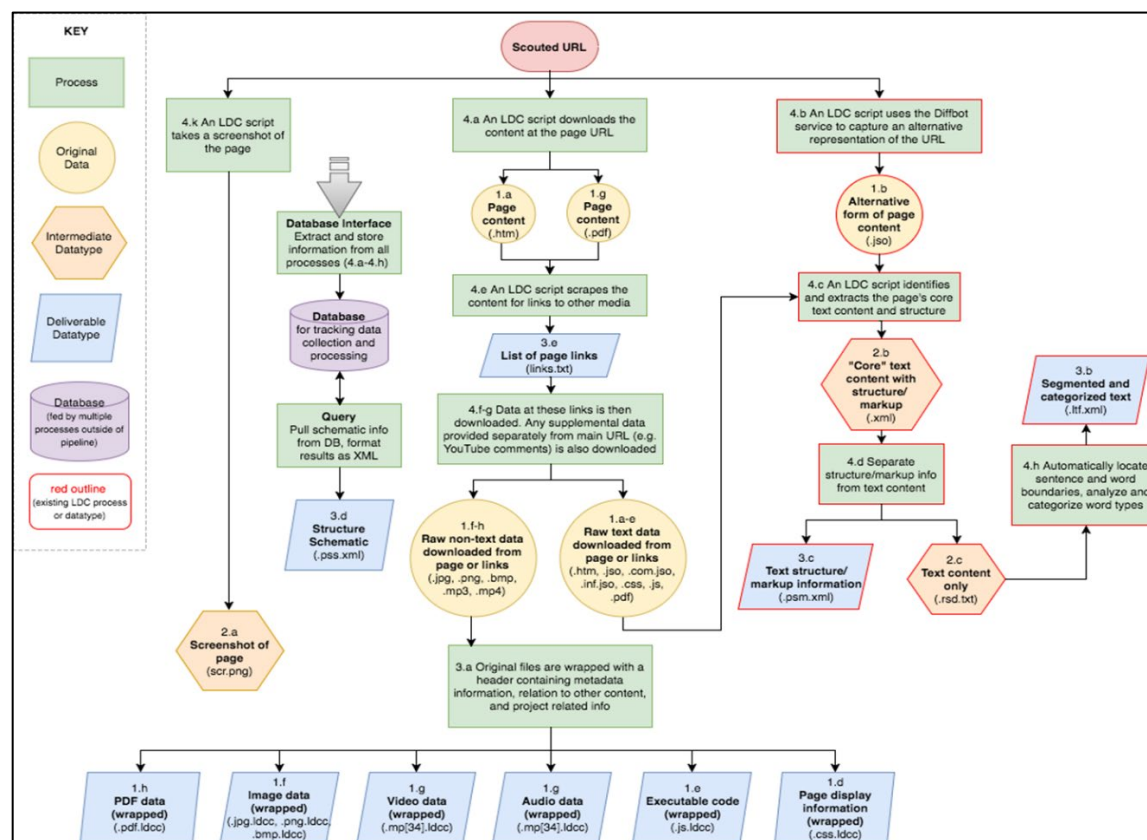


Figure 2: AIDA Multimedia Collection Pipeline

The collection process downloaded all text and non-text data from the target URL. Harvested HTML content was initially converted from its original form into a relatively uniform XML format; this stage of conversion eliminated irrelevant content (menus, ads, headers, footers, etc.), and placed the content of interest into a simplified, consistent markup structure.

The "homogenized" XML format then served as input for the creation of a reference "raw source data" (RSD) plain text form of the web page content; at this stage, the text was also conditioned to normalize white-space characters, and to apply other character/encoding normalization, as appropriate to the given language. This processing creates two xml files for each harvested "root" web page: psm.xml and ltf.xml. Primary Structural Markup (PSM) files contain information about the basic structure of the webpage, while Logical Text Format (LTF) presents a processed version of the text content of the page derived from the RSD.

The "homogenized" XML format described above preserves the minimum set of tags needed to represent the structure of the relevant text as seen by the human web-page reader. When the text content of the XML file is extracted to create the RSD format (which contains no markup at all), the markup structure is preserved in a separate "primary source markup" (psm.xml) file, which enumerates the structural tags in a uniform way, and indicates, by means of character offsets into the RSD file, the spans of text contained within each structural markup element.

For example, in a discussion-forum or web-log page, there would be a division of content into the discrete "posts" that make up the given thread, along with "quote" regions and paragraph breaks within each post. After the HTML has been reduced to uniform XML, and the tags and text of the latter format have been separated, information about each structural tag is kept in a psm.xml file, preserving the type of each relevant structural element, along with its essential attributes ("post_author", "date_time", etc.), and the character offsets of the text span comprising its content in the corresponding rsd.txt file.

The LTF data format is derived from RSD and contains a fully segmented and tokenized version of the text content for a given web page. Segments (sentences) and tokens (words) are marked off by XML tags (SEG and TOKEN), with "id" and character offset attributes relative to the corresponding rsd.txt file; TOKEN tags have additional attributes to describe the nature of the given word token (e.g., word or punctuation).

To the fullest extent possible, all discrete resources referenced by a given "root" HTML page (style sheets, JavaScript, images, media files, etc.) are stored as separate files of the given data type, and assigned separate 9-character file-IDs (the same form of ID as is used for the "root" HTML page).

In order to present these attached resources in a stable and consistent way, the LDC has developed a "wrapper" or "container" file format, which presents the original data as-is, together with a specialized header block prepended to the data. The header block provides metadata about the file contents, including the MD5 checksum (for self-validation), the data type and byte count, URL, and citations of source-ID and parent (HTML) file-ID.

An additional step was required for video files. In order to support NIST's evaluation design, LDC ran an implementation of the Cineast (Rossetto, et al., 2014) video segmentation software that was provided by NIST to produce a set of shot boundaries for each video file. This shot boundary

information was included with the source data so that human reference annotations and system output for video mentions of entities, relations, and events could refer to a shared set of keyframes.

Each corpus containing source data also contains documentation relating the various text, image, and video elements of a document (children) to the original web page (parent), along with additional information such as download date, publication date of the content (if known), manual and automatic language identification information, URL, and md5sums.

The results of the collection process for each Phase of AIDA are discussed in section 5.2 below.

3.3 Ontology/Tagset Development

The set of tags used to label entity, event, and relation types evolved significantly over the course of the AIDA program. The AIDA annotation tagset started with the adoption of an existing ontology developed for the prior DARPA DEFT program and was enriched for Phase 1 to reflect system and evaluation requirements. The existing DEFT ontology and corresponding annotation tagset was used for the first part of Phase 1, leading up to the Month 9 Pilot Evaluation, but it included a number of crucial gaps that needed to be filled in order to capture important informational conflict in the domain of Russia-Ukraine relations. These gaps were addressed for the final Phase 1 Evaluation by incorporating new domain-relevant tags as well as additional broadly relevant types proposed by the AIDA Program Ontology Working Group. Table 1 below shows the number of types annotated for the Pilot Evaluation and the final Phase 1 Evaluation.

Table 1: AIDA Tagset Expansion for Phase 1

	AIDA Pilot	AIDA Phase 1
Entity types	20	187
Relation types	23	49
Event types	47	139
Total types	90	375

The motivation for this significant expansion of the ontology was two-fold: it was intended to cover specific areas of informational conflict that were absent from the pilot tagset, but also to provide finer grained subtypes and sub-subtypes that AIDA systems might be able to utilize in distinguishing different hypotheses. For example, while the pilot tagset included a single conflict.attack event type, the final Phase 1 tagset broke this into 11 different sub-subtypes, distinguishing firearm attacks, bombings, and missile strikes among other sub-subtypes of attacks. Every event and relation in Phase 1 also included an “unspecified” sub-subtype that annotators could use as a backoff category when the more detailed type of attack (or other event/relation) was not clear from the data, or when the detailed sub-subtype was not present in the tagset.

The work of expanding the tagset was shared by LDC and performer teams, who organized an Ontology Working Group. In Phase 1, this working group has several sub-committees tasked with suggesting needed types for different kinds of annotation objects (one subcommittee on events, another on relations, another on entities, etc.). The task of consolidating these sometimes

incompatible components was undertaken by LDC. We standardized the depth of the ontology to three layers (type, subtype, and sub-subtype) from the one- to seven-level depths of the various components provided by the working group. We also eliminated easily confusable or ambiguous types, provided full definitions for types that were lacking them, standardized or provided argument role labels and type constraints on what kinds of entities could fill the argument slots, and processed the consolidated ontology into a human readable format for distribution to the AIDA community and a machine readable form for ingestion into our annotation infrastructure.

With each new scenario, additional effort was required to add new types that were relevant to the specific points of informational conflict identified in the topic development process. The annotation tagset for AIDA Phase 2 took the Phase 1 tagset as a starting point. In Phase 2, LDC worked with the AIDA Program Ontology Working Group to implement changes to allow events to be annotated as arguments of other events and to include topic arguments for contact events, in addition to customizing the tagset to capture important informational conflict in the Venezuela domain. This resulted in the addition of a number of new tags, but also the removal of some annotation types as subtypes were re-configured to accommodate the new types without creating overlapping categories. As part of this update to the tagset, 22% of the event types had their argument structure redefined to allow them to take events as arguments (e.g., He was arrested for shooting at parade attendees). The table below lists the number of types in Phase 1 and in Phase 2 and shows that while the overall number of types remained quite similar, the most significant change was in the number of events allowing other events as arguments.

Table 2: AIDA Tagset Expansion for Phase 2

	AIDA Phase 1	AIDA Phase 2
Event types	139	149
<i>Event types with event arguments</i>	0	33
Relation types	49	50
<i>Relation types with event arguments</i>	12	13
Entity/filler types	187	179

In Phase 3, the program made a dramatic shift toward a much larger and more open ontology by adopting Wikidata (Vrandečić and Krötzsch, 2014) to define entity, relation, and event types as well as to serve as the knowledge base for the program. A snapshot of Wikidata with certain kinds of duplication and unneeded areas removed was created by a team of AIDA researchers and used program-wide in order to provide a stable resource. This stable version of Wikidata is referred to as “DARPA Wikidata” or “DWD”.

In order to support the transition to DWD, LDC worked with the Ontology Working Group to merge the distinct but related annotation tagset that had been previously used in another DARPA program, KAIROS (DARPA, 2019) with the Phase 2 AIDA tagset to create a single, unified and generalized tagset based in DWD space, in place of the two highly customized, project-specific tagsets. Any new concepts added to the annotation tagset for the COVID-19 domain were based directly on DWD Q nodes. For example, the event type *SocialBehavior.Clean.Wash* was a new addition to the annotation tagset for Phase 3, and was based directly on the Q node Q23841 (washing - method of cleaning using flowing liquid).

While systems were encouraged to use any appropriate type they could find in Wikidata, exhaustive human annotation of multimedia data (required for TA1 evaluation and practice data) with a vast tagset of many thousand possible “types” was infeasible. Instead, LDC worked with the Ontology Working Group to both develop new types for the COVID-19 scenario and to map each of the LDC types to relevant Wikidata Q nodes. This allowed LDC annotators to work within a tagset similar in size to the previous phase while enabling systems to use the wider set of concepts in Wikidata. As will be detailed in the annotation sections below, LDC also made use of the full set of Wikidata types to provide additional type information on entities annotated with the LDC tagset, and also for the entirety of the TA3 Claim Frame annotation task.

3.4 Within-document Annotation

The purpose of AIDA within-document annotation was primarily to support the development and evaluation of TA1 systems performing extraction of entities, relations, and events that are then passed to TA2 systems for aggregation and representation in a common semantic representation within a cross-document knowledge base, which can then be used by TA3 to provide alternative hypotheses about events relevant to the scenario. LDC was tasked with providing a set of practice documents containing examples of the labeled events and relations necessary to characterize the various hypotheses present in the data, and to produce a set of reference data that could be used to measure system performance against blind evaluation topics. Thus, while the within-document annotation’s direct consumers were TA1 teams, it was necessary to consider the issues of topic salience and support for conflicting information in order for the data to be useful for TA2 and TA3 as well.

There were several changes to the within-document annotation approach across the phases of AIDA to support evolving program needs. The primary driver of those changes was the expansion of the tagset for entity, relation and event types discussed in section 4.3 above. Another driver of changes in the within-document annotation approach was the push toward increasingly exhaustive annotation of entity, relation, and event mentions within the documents selected for annotation to support both TA1 performer requests and NIST evaluation design. The push toward larger tagsets and simultaneously more exhaustive mention annotation required creative approaches to data selection and scoping of annotation to prevent TA1 annotation from completely overtaking the available annotation resources. The following sections will describe the basic structure of entity, relation, and event annotation, which was stable across the phases of the program, followed by a discussion of the changes made to the scope of annotation to maintain the feasibility of producing high quality data in the context of a growing tagset and the need to support exhaustive annotation.

3.4.1 Basic Structure and Approach

In all phases of annotation, the basic structure of within-document annotation remained the same: labeled mentions of entities, relations, and events (including argument structure for events and relations) within the multimedia document. Within-document annotation also included coreference of mentions within the same document. All annotation was performed by trained annotators who had native fluency in the language of the document.

Using information collected during data scouting, documents known to contain query-responsive information were selected for manual annotation.

For each event or relation subject to annotation, AIDA annotators made a number of decisions. First, each event or relation, and each associated entity or filler argument, was anchored in document-level provenance. Provenance for text instances was recorded as character offsets. In the pilot annotation effort, provenance for image and video instances consisted of a document ID only. In the Phase 1 annotation, image instances were labeled with bounding box provenance, while video instances included keyframe IDs, bounding boxes and/or timestamps as appropriate. In Phase 2 and Phase 3, bounding box annotation was limited to entity mentions, while event and relations were anchored in the whole image. Annotators provided a brief text description (a word or short phrase) for each event, relation or argument and assigned it a type from the annotation tagset. Arguments were also labeled for the role they play in the event or relation, e.g., the perpetrator vs. victim of an attack event. Text instances of arguments were also labeled as name, nominal or pronoun.

Annotators then specified any attributes associated with the event, relation, or argument. Two attributes used throughout all phases were “not”, indicating negation, and “hedged”, indicating uncertainty. “Not” is used to indicate that the source asserts that the event or relation did not happen, or that the argument did not participate in the event. “Hedged” is used to indicate uncertainty as reported by the source. A mention can combine “hedged” and “not” attributes, indicating that that source asserted that the relation or event possibly or likely did not happen. In Phase 3, the additional attributes “generic” and “irrealis” were added. The “generic” attribute was used in order to be able to capture claims were general assertions about how the world works, as opposed to specific instances of and event (e.g. “Wearing masks (in general) prevent the spread of COVID” vs “Wearing a mask prevented me from getting COVID when I attended an in-person work meeting.”). The “irrealis” attribute was used to capture claims about things that involve wishes/hopes, counterfactuals, etc. (e.g., “If the US had closed the border sooner, COVID would not have spread to this country.”).

In the pilot annotation, events could be labeled as “deliberate” or “accidental”, capturing explicit assertions by the source about whether the event was intentional or not. Some events could further be labeled as “legitimate” or “illegitimate”, to capture political legitimacy of elections. In Phase 1, these attributes were replaced with new relation types which captured the source of the assertion about deliberateness or legitimacy.

Finally, relations and events were labeled for temporal information. Annotators specified start and/or end dates for the relation or event when that information is present in the document, characterizing the information with as much precision as possible given document information. Dates are characterized as starting (ending) on, before or after a particular date, and the date is

expressed in year-month-date format, with partially populated dates possible. For instance, a timestamp of *start on 2015-01-XX* indicates that the event or relation started sometime on or after January 2015, with a more precise date unknown. If the document does not contain any date information for the event or relation, annotators choose a start and/or end date of unknown.

For both Phase 2 and Phase 3, the TA1 annotation pipeline began with two separate branches of entity annotation: first-pass annotation of text entity mentions by LDC, and first-pass annotation of image mentions by PAR on a common set of source documents. These first-pass LDC and PAR entity annotations were then consolidated into a single set of entity mentions in a uniform, machine-readable format, which was then imported into LDC's annotation platform and underwent a quality-control process before being passed as input to the event and relation stage of annotation. This process allowed the annotation team to make faster progress through the laborious process of exhaustively tagging entities and supplied a stable, quality-checked set of entity mentions that were then available for the event and relation annotators to select as fillers of argument roles.

3.4.2 Changes in Scope and Exhaustiveness

In Phase 1, annotation proceeded in two stages. To support the AIDA program pilot evaluation, which was intended to help program performers define fundamental concepts and to establish baseline system performance, LDC produced a set of annotated pilot topics. The larger Phase 1 corpus included the official practice and evaluation topics for the final phase evaluation.

AIDA pilot and Phase 1 annotation consisted of labeling only topic-relevant events and relations, along with their arguments and attributes. This meant that mentions of events and relations that were present in the document, and which met the definition of an available type in the tagset were annotated only if they were salient to the scenario and topics of interest, and mentions of entities and fillers were annotated only if the entity or filler participated as an argument in a relevant event or relation. This approach was consistent with the requirements of the BAA that "the 'ground truth' will consist of all knowledge elements relevant to the various hypotheses" (DARPA, 2017, p.15).

For each document, annotators labeled events and relations relevant to a single topic, ignoring information about other topics even when it was present in the document. Because annotation was restricted to topic-relevant information only, it was necessary to create an operational definition of relevance, or "topic salience", that would strike the right balance between completeness and precision. The definition had to balance the need to constrain the inclusion of irrelevant or marginally relevant events and relations in the annotation against the need to allow for the presence of unexpected information in the document, which could lead to discovery of a novel hypothesis about the topic. In addition, the definition of salience needed to be intuitive and promote consistent annotation.

In the pilot effort, topic salience was initially defined as "critical to understanding the topic, and/or frequently associated with the topic." Although annotators found this definition to be reasonably intuitive, it led to over-annotation of irrelevant or tangentially relevant events and relations. The definition was then revised to "supports one or more facet-level hypotheses," (where a facet-level hypothesis was defined as an answer to one of the queries in the topic model), but in practice this definition was too constrained, and prohibited annotation of relevant items that were needed to support construction of semantically coherent topic-level hypotheses. For instance, information about affiliations between entities or supporting events such as the transfer of a weapon from one

entity to another were often left out when the annotation was restricted to supporting the hypothesis that a particular entity fired on protesters. This restrictive definition also precluded discovery of new hypotheses that were not part of the a priori set constructed during topic development.

Based on the lessons learned during pilot annotation, for the final Phase 1 evaluation, salience was redefined as “relevant to one or more queries.” By focusing on relevance to the queries rather than to the hypotheses, this working definition of topic salience sufficiently excluded the presence of irrelevant or marginally-relevant information in the annotations, while allowing annotators to capture the full set of events and relations needed to represent topic-level hypotheses; furthermore it did not preclude annotators from labeling information about potential novel hypotheses that were query-responsive.

Because the goal of AIDA annotation was to provide examples (in the case of practice topics) and reference annotations (in the case of evaluation topics) for the events and relations (plus their arguments) that were required for hypothesis construction, it was not necessary to label every occurrence (i.e. mention) of a given event or relation in the data. Furthermore, by program design the annotated data was not intended to serve as training data for basic information extraction, where exhaustive annotation of all event and relation mentions may be required. Given this context, a pragmatic decision was made to limit annotation to a single occurrence of each event or relation (plus arguments) per each kind of document element, where document elements are the set of individual text, image and video files that make up a complete “parent” document. For example, if the same event is depicted in both the text and an image of a document, it is labeled once in the text and once in the image; if the same event appears 3 separate times in the video portion of the document, it is labeled only once for the video document element.

At the request of AIDA system developers, some additional event and relation instances were labeled in Phase 1, expanding on the approach taken in the pilot. First, audio and video tracks of video document elements were treated as distinct for purposes of annotation, such that if an event appeared in video with both audio and visual evidence for the event, annotators would label one event in the audio track, and a separate event in the video track. Second, arguments that occurred in closest proximity to the “trigger” (word or keyframe) where the event or relation first appears in a document were always labeled as “base” mentions, with more distant argument mentions being optionally labeled as additional “informative” mentions. Third, as time permitted annotators labeled additional informative mentions of events/relations and their arguments beyond the single instance per document element, with a strong preference for instances that provided especially helpful instantiations of the event, relation or argument - for instance, a very clear video image or a detailed text description naming the actor in some event. Finally, a handful of documents were exhaustively labeled with all salient events and relations plus their arguments and attributes, to provide a resource for system error analysis.

In Phase 2, a shift in the evaluation paradigm from assessment-based scoring to automatic scoring resulted in a shift from salience-based annotation to exhaustive annotation as required by NIST’s scoring approach. Fully exhaustive annotation of the 378 types in the annotation ontology across every piece of media in every document, including images and video, without limiting mentions to those salient to a specific topic, would have required enormous amounts of time and resources and would have been counter to DARPA’s direction to keep as much focus as possible on the hypothesis

creation aspect of the program rather than the diagnostic measure of TA1 extraction performance. The annotation method was consequently reimagined to allow annotators to tag information that supported automatic scoring in each document without risking the explosiveness that would naturally result from fully exhaustive annotation of entire documents with the full tagset.

Narrowing the scope of exhaustive annotation in Phase 2 began with narrowing what types of information would be annotated and how much of each document would be annotated. As a preliminary step before annotation began, annotators identified salient events and relations in the corpus documents using the Phase 1 definition of “salient” as “relevant to one or more queries.” Annotators then identified regions of each document (up to three paragraphs of text or 30 seconds of video, plus whole images) that contained a mix of both salient and non-salient events and relations. The result of this auditing pass over the documents was a collection of annotatable regions of the documents that were guaranteed to contain salient events and relations, and that could be annotated exhaustively within the level of effort LDC was directed to allocate to this diagnostic measure.

In the annotation pass, events and relations were only tagged in the regions of the documents selected in the previous auditing pass. Each document was assigned a closed list of event and relation types based on the previous auditing pass. Only events and relations of those types could be tagged within each document region, and mentions of events and relations of the selected types within the annotation regions were tagged exhaustively. We then found all the entity types annotators used to fill the arguments of the events and relations they had tagged in these regions and created another closed list of entity types that could be tagged in each document. In a final pass, annotators exhaustively tagged mentions of the selected entity types throughout each document. The annotation of each document therefore includes exhaustively tagged events and relations of types known to be salient to a topic tagged in manageable document regions that are also known to be salient, as well as exhaustively tagged entities throughout each document that match the types of entities known to have participated in salient events and relations.

In Phase 3, the need for exhaustive annotation to support TA1 diagnostic evaluation remained, and NIST and performers both indicated a preference for full documents annotated exhaustively with the same set of tags across all documents. The same set of issues surrounding the feasibility of exhaustive annotation with a large tagset remained, so another approach to scoping was required. The resulting approach was to exclude from annotation certain types which, especially in images, can create an explosion of annotations when not constrained by the filter of topic-salience. An example of such a type would be “Body Part”, since it could lead to the annotation of any and all people present in images with multiple bounding boxes around different body parts. Figure 3 below illustrates the number of mentions that might be created when an image is annotated only for entities, relations, and events that are salient to the topic, while Figure 4 shows the explosion of irrelevant annotations that occurs when there is no salience constraint.

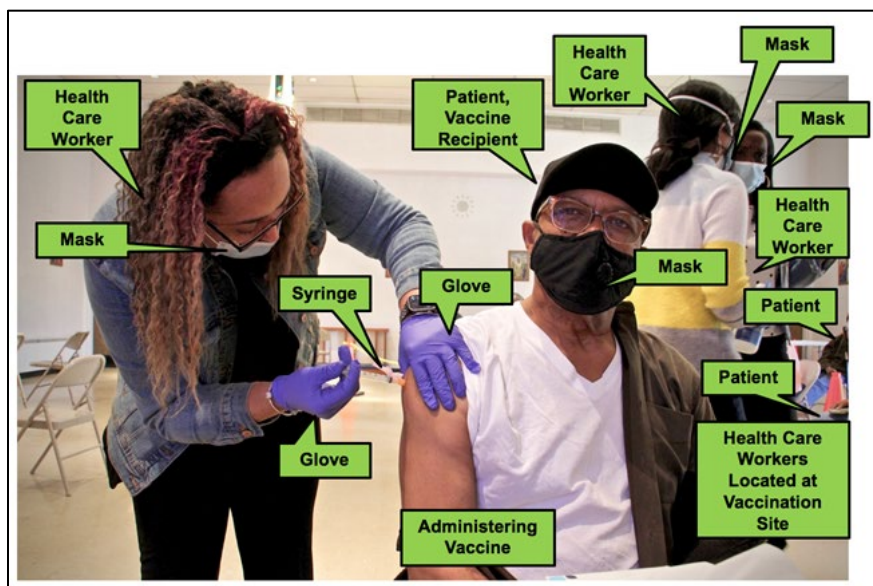


Figure 3: Saliency-Constrained Image Annotation Example

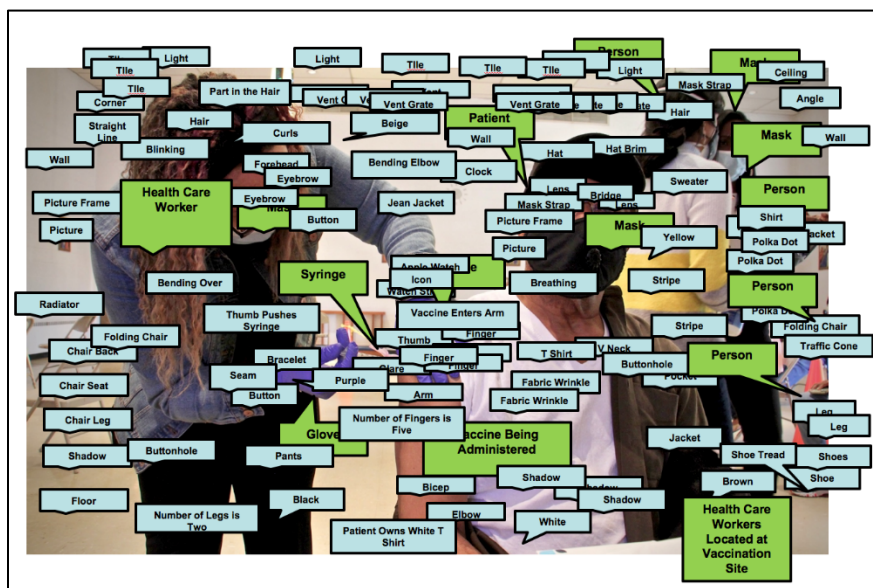


Figure 4: Unconstrained Exhaustive Image Annotation Example

The results of within-document annotation for all three phases are discussed in section 5.4.

3.5 Cross-document Annotation

Cross-document annotation in AIDA was relatively stable across the program's phases, with the major changes being related to the reference KB used for the KB linking task.

For all events, relations and their arguments labeled in the data, annotators performed coreference. To support creation of corpus-wide hypotheses, cross-document coreference was a requirement; for AIDA, cross-document necessarily means both cross-language and cross-media. Procedurally, coreference was achieved by manually linking individual entity and event instances to a knowledge base (KB).

For the pilot annotation effort, LDC created individual "mini-KBs" for each topic, containing only the events, relations and entities (arguments) anticipated to be relevant for that topic. Annotators manually linked individual event, relation and entity instances to the KB and flagged any items that could not be linked so that new KB entries could be created.

The Phase 1 evaluation design required a program-wide reference entity KB, constructed by LDC to include salient entities identified during topic development as well as a large number of general domain entities drawn from the LORELEI reference KB (Tracey and Strassel 2020). Coreference annotation for entities consisted of manually linking entity instances to the reference KB. When no match is present in the KB, the entity is marked as NIL; once all KB linking is complete, all NILs are reviewed and clustered, such that multiple mentions of the same NIL entity are assigned a unique NIL ID. Events were also manually clustered and assigned unique NIL IDs. Finally, relations were automatically clustered and assigned unique NIL IDs based on the results of manual entity clustering: relations with the same type, and whose arguments have the same argument role and contain the same entity (KB or NIL) ID, are considered coreferential. Phase 2 used the same KB as Phase 1, with the addition of a number of entities specifically relevant to the Phase 2 Venezuela Scenario. The Phase 2 KB was only used in assessment, as KB linking annotation was not needed for the evaluation reference annotation and so was dropped in favor of more exhaustive within-document annotation and more complex TA3 assessment. While linking to the reference KB was not performed, within-document coreference NIL clusters was provided.

In Phase 3, the DWD served as the KB. This meant that the KB was very large, but that there was no manual addition of entities known to be relevant to the COVID-19 scenario. While the KB changed for Phase 3, the overall approach to coreference and KB linking did not. Annotators provide KB in the form of a DWD Q Node that points to the entry in DWD that identifies the exact individual entity represented by the mentions; in the event that no appropriate Q node identity link is found, annotators assign a NIL ID as in previous phases.

While the approach to KB linking and coreference remained the same, some additional information about the entities was recorded by annotators during the KB linking task for Phase 3, once they had created the entity clusters and linked them or assigned NIL IDs.

In the Coreference and KB Linking step of annotation, annotators assigned a base type to entity clusters. The base type is a top-level type (i.e., not including subtype or sub-subtype) selected from the annotation ontology and is intended to be the best available representation of the entity's real-

world type independent from in-document usage. In most cases, an entity’s base type matches the top-level annotation ontology type assigned to mentions of that entity in the document, but the base type can differ when in-document usage is at odds with the type that would best represent the entity outside of the context of the document.

Table 3: Entity Base Types Assigned in KB Linking

Cluster 1: Barack Obama		Cluster 2: Robinhood Markets	
Mention 1 Obama	Annotated type Type: PER Subtype: Politician Sub-subtype: HeadOfGovernment	Mention 1 Robinhood	Annotated type Type: ORG Subtype: CommercialOrganization Sub-subtype: Corporation
Mention 2 the former president	Annotated type Type: PER Subtype: Politician Sub-subtype: HeadOfGovernment	Mention 2 the Robinhood team	Annotated type Type: PER Subtype: ProfessionalPosition Sub-subtype: Unspecified
Mention 3 a senator	Annotated type Type: PER Subtype: Politician Sub-subtype: Unspecified	Mention 3 them	Annotated type Type: ORG Subtype: CommercialOrganization Sub-subtype: Corporation
Base type assigned to cluster PER		Base type assigned to cluster ORG	

In the example in Table 3, the base type of Cluster 1 matches the base type of all annotated mentions of the given entity, while the base type of Cluster 2 best represents the entity’s real-world type despite being at odds with the annotated type of one mention in the cluster. While the PER “team” mention may appear to be a different entity from the ORG mentions in isolation, the document context makes clear that all of these mentions refer to the same entity, and thus the cross-type coreference cluster is allowable; it is precisely this kind of potential difference between the interpretation of mentions in isolation vs. mentions in context that gave rise to the addition of the base type information to the clusters created in the KB Linking step of annotation.

In addition to the base type information, annotators also assign a fine-grained type link in the form of a DWD Q Node for entities. Annotators are not confined to the entity types available in the annotation ontology for this part of the task. Instead, annotators are instructed to search DWD for the entry that most closely corresponds to the real-world type of the entity they are tagging regardless of the DWD entry’s correspondence to the AIDA annotation ontology types. Type Q Nodes are meant to point to the most specific type possible for the entity. The most accurate and

specific real-world type for an entity might vary by in-document usage, so it is possible for any given entity to have several different correct fine-grained type links.

Table 4 below shows examples of several entity clusters with all of the additional information provided in the Phase 3 task.

Table 4: Examples of Entity Coreference Clusters

Entity Mention	Annotated Mention Type	Coref Cluster	Base Type	Identity KB Link	Type KB Link	Generic Status
Anthony Fauci	scientist (Q901)	1	person (Q215627)	Anthony Fauci (Q573246)	immunologist (Q15634285)	FALSE
Timothy Barrett	spokesperson (Q17221)	2	person (Q215627)	NIL_E11111	spokesperson (Q17221)	FALSE
Wuhan Institute of Virology	organization (Q43229)	3	organization (Q43229)	Wuhan Institute of Virology (Q30292611)	research institute (Q31855)	FALSE
the Wuhan lab	building (Q41176)					
Advil	medical equipment (Q6657015)	4	artificial physical object (Q8205328)	NIL_E99999	Advil (Q16240691)	TRUE

3.6 Hypotheses

In the Phase 1 pilot effort, annotation included explicit annotation of query-level hypotheses (i.e., the answers to the facet-level queries included in the topic model), but in the final Phase 1 evaluation and in Phase 2, the emphasis was shifted to the larger topic-level hypotheses. And in Phase 3, the emphasis shifted again to claim frames.

3.6.1 Phase 1: Facet-level hypotheses and prevailing theories

In the pilot hypothesis annotation, for each relation or event labeled in a document, annotators indicated whether and how that item was relevant to the set of hypotheses for that document's

topic, specifying whether the event/relation fully or partially supports the hypothesis or contradicts the hypothesis. Each event or relation was marked as relevant to zero, one, or more than one hypothesis. For example, one of the facet-level hypotheses for the query “Who fired on protesters and/or police at the Maidan protests?” is “Members of the Berkut, a police force loyal to the Yanukovych government, fired on protesters.” An attack event taking place at Maidan with *Berkut* as the attacker and *protesters* as the victim would be marked as fully supporting the hypothesis. In contrast, an attack event with *security forces* as the attacker and *guns* as the weapon would be marked as partially supporting the hypothesis. An attack event in which *security forces* were the **victim** would be marked as contradicting the hypothesis.

Figure 5 shows an example of two facet-level hypotheses for a single query.

Query: Who fired on protesters and/or police at the Maidan protests?

H1: Members of the Berkut, a police force loyal to the Yanukovych government, fired on protesters.

H2: Snipers loosely affiliated with the Ukrainian government fired on protesters.

Figure 5: Query and Facet-Level Hypotheses

In addition to facet-level hypotheses, more complex topic-level hypotheses were also needed to support the evaluation, since AIDA systems must aggregate information from across the whole corpus to produce complex and semantically coherent knowledge graphs that present different hypotheses about what happened with respect to the topic as a whole. Reference topic-level hypotheses took the form of “prevailing theories”, which describe domain expert expectations about the differing accounts or perspectives on the topic that are likely to occur in the data based on prior knowledge of the scenario. While facet-level hypotheses address a single information element, prevailing theories are often responsive to multiple queries and represent a coherent perspective within the larger topic narrative. Prevailing theories were largely informed by MITRE’s scenario documentation, but also shaped by LDC’s preliminary annotation of the corpus to provide some insight into which of the potential prevailing theories actually appeared in the data.

For each prevailing theory identified, we created a natural language description characterizing this account of the topic, plus a list of the events and relations, along with their arguments (entities), that would be required as part of a knowledge graph that adequately reflects this theory. Some prevailing theories with more narrow scope include only a short list of required events and relations, while some with broader scope include a much longer list.

Figure 6 illustrates one natural language prevailing theory along with the corresponding list of events and relations that would be required in a system hypothesis that fully captures the content of the prevailing theory.

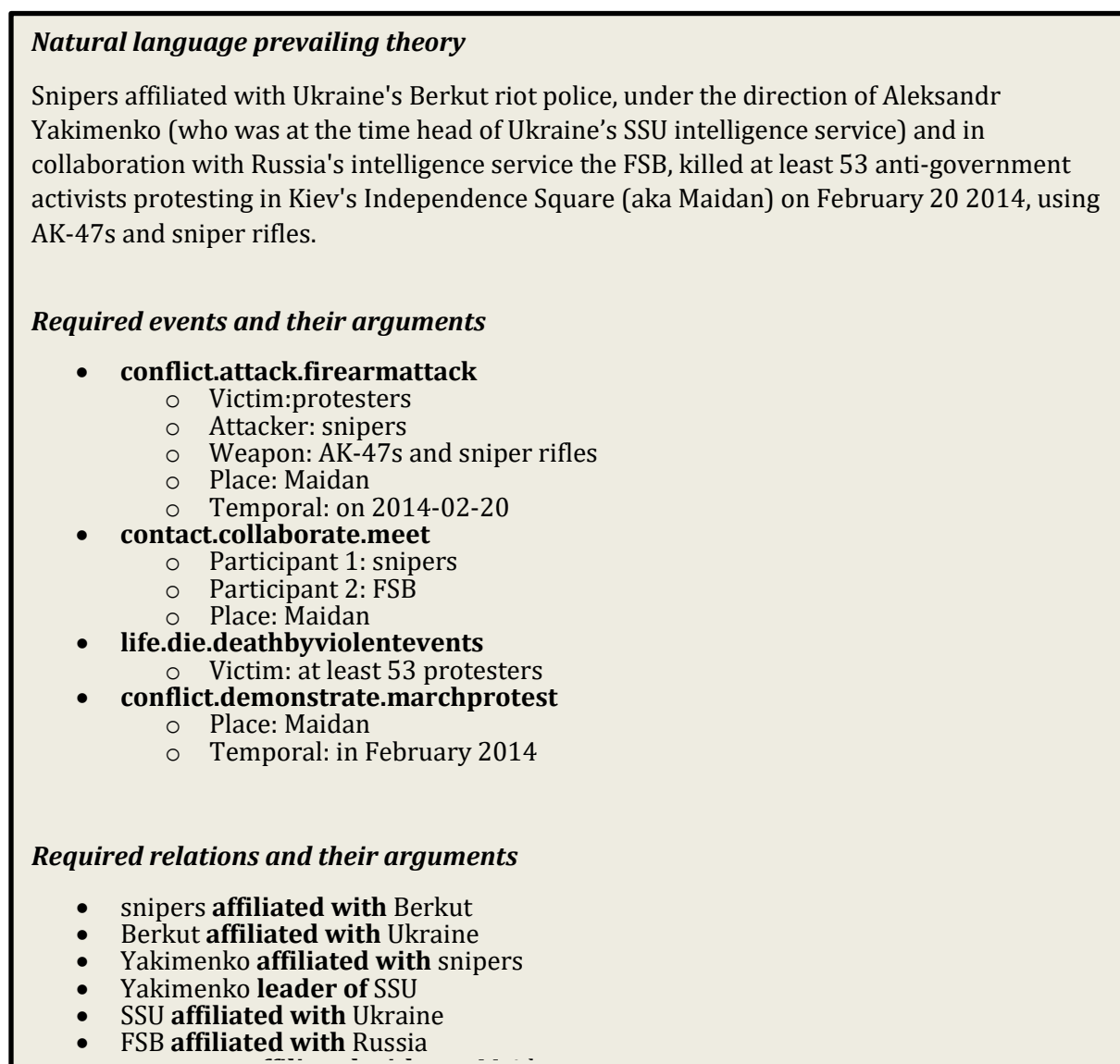


Figure 6: Example Prevailing Theory with Events and Relations

Given the design for the final Phase 1 evaluation, direct annotation of facet-level hypotheses was no longer necessary. Instead, the emphasis shifted to development of the prevailing theories characterizing the broader topic-level hypotheses expected to be present in the corpus. These prevailing theories were then used as part of the post-hoc assessment of system output, as discussed in section 4.8.

3.6.2 Phase 2: Perspectives

In Phase 2 we expanded beyond the prevailing theories from Phase 1 and developed full perspectives on the topics. Each perspective contained an understanding of the informational conflict in the topic as a whole from that perspective, along with a matrix of knowledge elements (events or relations and their arguments with minimal impoverished structure, no types, and no argument roles) that comprised the understanding in the perspective. This provided a more complete representation of the perspective on the topic's informational conflict, without requiring formal, detailed representation of all information in the topic.

The final step in creating each perspective was to add a "KE matrix" to support assessment and scoring. The matrix included each event or relation in the perspective, along with its arguments. The KEs in the matrices did not include tagset types, because the evaluation and manual assessment did not require exact type matching. However, an assumed KE type helped to determine the set of arguments for each event or relation in the matrix. The intention was to recognize that there might be more than one legitimate way to represent the same set of 'facts' that are relevant to the perspective in the data using the tagset, and not including types in the matrices avoided penalizing systems for choosing a different, but still logical, way of representing the information.

Example Perspective:

- Summary: Opposition's Perspective: Miguel Castillo was killed by government security forces with a marble
- Background Information: In 2017, anti-government protesters held demonstrations and marches in Venezuela. Thousands of people were injured in the 2017 protests in Venezuela. Over 100 people were killed in the 2017 protests in Venezuela. The anti-government demonstration, "Our Shield is the Constitution", was held on May 10, 2017.
- Details: Miguel Castillo was killed by security forces (an example being National Guardsmen -- aka Bolivarian National Guard (GNB)) on May 10 from being shot with a marble used as a bullet.
- KE Matrix: See Figure 7

Matrix KE Item Number	Natural Language KE Description	Item KE	Required vs. Optional	Coref	"OR" Groups	2-EVENT	Perspective Element(s)
KE001	In 2017, anti-government protesters held demonstrations and marches	event	opt				BG_E202_S001
KE001-x1	anti-government protesters	entity arg	opt	coref1			BG_E202_S001
KE001-x2	Venezuela	entity arg	opt				BG_E202_S001
KE002	Affiliation relation between protesters and anti-government side	relation	opt				BG_E202_S001
KE002-x1	anti-government protesters	entity arg	opt	coref1			BG_E202_S001
KE002-x2	anti-government	entity arg	opt				BG_E202_S001
KE003	Thousands of people were injured in the 2017 protests in Venezuela	event	opt			x2 (injure,attack)	BG_E202_S002
KE003-x1	Thousands of people who were injured	entity arg	opt	coref2		x2 (injure,attack)	BG_E202_S002
KE003-x2	Venezuela	entity arg	opt			x2 (injure,attack)	BG_E202_S002
KE004	Over 100 people were killed in the 2017 protests in Venezuela	event	opt			x2 (die, attack)	BG_E202_S003
KE004-x1	Over 100 people who were killed	entity arg	opt	coref3		x2 (die, attack)	BG_E202_S003
KE004-x2	Venezuela	entity arg	opt			x2 (die, attack)	BG_E202_S003
KE005	Count relation for the thousands of people who were injured	relation	opt				BG_E202_S002
KE005-x1	Thousands of people who were injured	entity arg	opt	coref2			BG_E202_S002
KE005-x2	thousands	entity arg	opt				BG_E202_S002
KE006	Count relation for over 100 people who were killed	relation	opt				BG_E202_S003
KE006-x1	Over 100 people who were killed	entity arg	opt	coref3			BG_E202_S003
KE006-x2	over 100	entity arg	opt				BG_E202_S003
KE007	The anti-government demonstration, "Our Shield is the Constitution" [Nu]	event	opt				BG_E202_S006
KE008	(National Guardsmen, security forces) shot and killed Miguel Castillo Br	event	req			x2 (die, attack)	E202_Q013, E202_Q014, BG_E202_S006
KE008-x1	Miguel Castillo Bracho	entity arg	req			x2 (die, attack)	E202_Q013
KE008-x2	National Guardsman	entity arg	req			x2 (die, attack)	E202_Q013
KE008-x3	security forces	entity arg	req		group1	x2 (die, attack)	E202_Q013
KE008-x4	marble	entity arg	req		group1	x2 (die, attack)	E202_Q014
KE008-x5	Venezuela	entity arg	opt				BG_E202_S002

Figure 7: Example KE Matrix

3.6.3 Phase 3: Claim Frames

In a large change to the concept of hypotheses to accompany the new COVID-19 domain and the proliferation of misinformation regarding the pandemic the focus of TA3 hypothesis annotation and assessment shifted from perspectives to claim frames. A completely new claim frame annotation task was added to the Phase 3 annotation pipeline.

In support of this new task, DARPA, MITRE, LDC, and AIDA teams engaged in a coordinated “hackathon” to come to a consensus on the evaluation and annotation goals for TA3. The hackathon efforts developed an approach to TA3 annotation that involved identifying claims in a document, identifying knowledge elements (entities, events, relations) that supported those claims, and linking the knowledge elements to nodes in DWD.

As the hackathon process took place over the course of several months, LDC produced iterative updated samples of manually annotated claim frames, as well as updated corresponding knowledge element annotations. Each successive version of the sample annotation reflected changes to the claim frame annotation principles as decided in the hackathon. LDC also coordinated AIDA teams’ parallel efforts to produce manual claim frame annotations, for comparison to LDC’s sample annotations.

After the hackathon ended, LDC coordinated AIDA teams’ continued efforts to produce manual claim frame annotations for particular sets of documents selected by LDC from the Phase 3 practice topic source data. LDC worked with each AIDA team to fix errors in the teams’ annotations and to prepare their annotations for inclusion in LDC data releases.

In the context of Claim Frame Annotation, a claim is a statement made implicitly or explicitly in a document that fills the X variable in a claim template (e.g., “X cures COVID-19”). In Claim Frame Annotation, annotators search documents for claims and, once they’ve found one, they label and identify information that contextualizes the claim being asserted. With the shift to claims for the Phase 3 evaluation, annotators were able to produce single-variable summaries of topic-specific informational conflict that included both natural language descriptions and supporting events and relations. While none of the prior phases attributed the narratives or perspectives to any particular entities, claim frames included extensive slot fills and opportunities to describe who was making a particular claim.

Based on high-level principles discussed in the hackathon, LDC developed a formal set of claim frame field definitions for the teams to use in the hackathons and in team TA3 annotations, as well as for use by LDC’s own TA3 annotations. The table below provides a summarized view of the claim frame format.

Table 5: AIDA Claim Frame Template

Field Name	Description
Document ID	root UID for the parent document containing this frame
Claim ID	unique identifier for claim frame
Topic	claim topic, drawn from official list (see table XXX)
Subtopic	claim subtopic, drawn from official list (see table XXX)
Claim Template	claim template, drawn from official list (see table XXX)
X Variable	value of the X variable in the claim template
X Variable Identity Q Node	identity Q Node for X variable
X Variable Type Q Node	type Q Node(s) for X variable
Natural Language Description	natural language summary of the claim captured in this frame
Claim Semantics	list of event and/or relation KE IDs that comprise the claim itself
Claimer	claimer identity
Claimer Identity Q Node	identity Q Node for claimer
Claimer Type Q Node	type Q Node(s) for claimer
Claimer Provenance	text extract from document justifying the claimer, where available
Claimer Affiliation	affiliation of claimer
Claimer Affiliation Identity Q Node	identity Q Node for claimer affiliation
Claimer Affiliation Type Q Node	type Q Node(s) for claimer affiliation
Epistemic Status	claimer's stance on truth and certainty of populated claim template
Sentiment Status	claimer's sentiment towards the populated claim template
Claim Date/Time	time/date when claim was made
Claim Location	location where claim was made
Claim Location Identity Q Node	identity Q Node for claim location

Claim Location Type Q Node	type Q Node(s) for claim location
Claim Location Provenance	text extract from document justifying the location, where available
Claim Medium	medium where claim was made
Claim Medium Identity Q Node	identity Q Node for claim medium
Claim Medium Type Q Node	type Q Node(s) for claim medium
Claim Medium Provenance	text extract from document justifying the medium, where available
Associated KEs	list of all event and relation KE IDs associated with this claim, including claim semantics

Though each document was selected for annotation based on its relevance to a single topic, information that fills the X variables in any number of scenario-relevant topics might appear in any document. All claims that appear in a document are annotated, but only claims designated as practice topics are annotated in the practice topic document set and only claims designated as evaluation topics are annotated in evaluation document set. Each unique claim – defined as a unique combination of the Topic/Subtopic, X Variable, Claimer, Epistemic Status, Sentiment Status, Claim Timestamp, Claim Location, and Claim Medium fields – is tagged only once per document. If any of the listed fields would be filled by a different piece of information, that justifies tagging a new claim, as illustrated in Figure 8.

Claim	Topic – Subtopic	X Variable	Claimer	Epistemic Status
Facebook posts have claimed that weed is a cure for coronavirus	Curing/ Preventing/ Destroying the Virus - Curing the virus	Cannabis	Facebook posters	True-Certain
Former NFL Player Kyle Turley Claims Cannabis Can Cure Coronavirus	Curing/ Preventing/ Destroying the Virus - Curing the virus	Cannabis	Kyle Turley	True-Certain

Figure 8: Unique Claims with Different Claimer

KEs (events and relations, as well as their arguments) are only to be tagged when they are needed to support a claim that is being asserted, or to contextualize information that is filling a claim field.

The type restrictions that apply to TA1 annotation do not apply to TA3 annotation. Instead, each filler is linked to both an Identity Q node in DWD and also a fine-grained type Q node in DWD.

Claims and their supporting or contextualizing information can be mentioned either explicitly or implicitly anywhere in the document. Annotators therefore do not supply fine-grained provenance in this task, and the annotation data does not include text offsets or image bounding box coordinates. As a result, each unique event, relation, and entity is tagged only once, and the entire document is considered the provenance of the entire claim. Table 6 below shows an example of a completed claim frame annotation.

Table 6: Example Completed Claim Frame Template

Claim 82					
Topic/Subtopic			Origin of the virus/Where the virus originated		
X Variable Description	United States	X Variable Identity Q Node	United States of America (Q30)	X Variable Type Q Node	country (Q6256)
Natural Language Description			Zhao Lijian promotes idea that virus might have originated in US		
Claim Semantics + Associated Mentions			List of supporting KEs		
Claimer Description	Zhao Lijian	Claimer Identity Q Node	Zhao Lijian (Q87084117)	Claimer Type Q Node	Foreign Ministry Spokesperson of the People's Republic of China (Q4528843)
Claimer Provenance			Zhao Lijian		
Epistemic Status		True-Uncertain	Sentiment Status		Neutral-Unknown
Claim Timestamp			On 03/12/2020		
Claim Medium Description	Twitter	Claim Medium Identity Q Node	Twitter (Q918)	Claim Medium Type Q Node	social networking service (Q3220391)

Claimer Affiliation (1) Description	China	Claimer Affiliation (1) Identity Q Node	China (Q148)	Claimer Affiliation (1) Type Q Node	country (Q6256)
Claimer Affiliation (2) Description	Chinese foreign ministry	Claimer Affiliation (2) Identity Q Node	Ministry of Foreign Affairs of the People's Republic of China (Q1155502)	Claimer Affiliation (2) Type Q Node	Constituent Department of the State Council (Q4294273)

Cross-Claim Relations annotation is the final stage in the Claim Frame annotation pipeline. In this stage, annotators compare each unique claim to each other unique claim that has been annotated in the corpus and indicate the epistemic relationship between each claim pair. In preparation for this step, the claims tagged in TA3 annotation are rendered as populated claim templates, which insert the Claimer, X Variable, and Epistemic Status fields of a claim into a modified version of the corresponding subtopic's claim template (see Figure 9 below). Doing so allows annotators to make efficient, common-sense decisions about the relationships between claims that focus on potential points of informational conflict. In an additional preparatory step, annotators cluster duplicate claims into equivalence classes to reduce the total number of claims for comparison. Two claims are considered duplicates of each other if they share a Claimer, X Variable, and Epistemic Status. Figure 10 shows the merging of duplicate claims in which the claims from two different documents, each asserting that social media users claim that the virus originated in a Wuhan Institute of Virology lab, are merged into a single equivalence class for comparison to other claims.

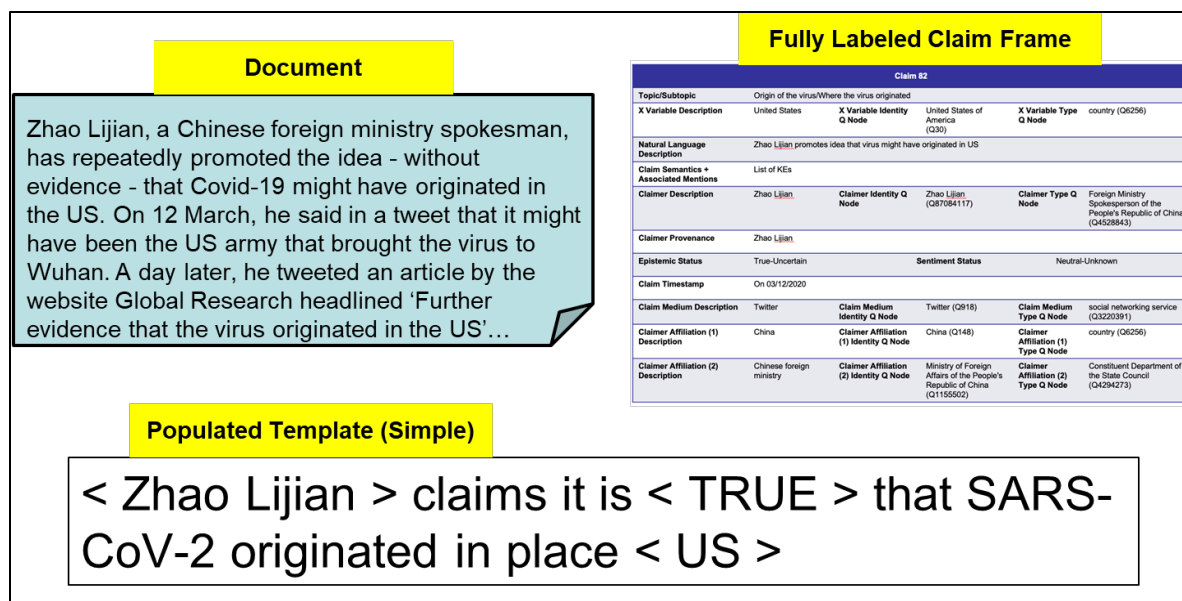


Figure 9: Populated Claim Template Example

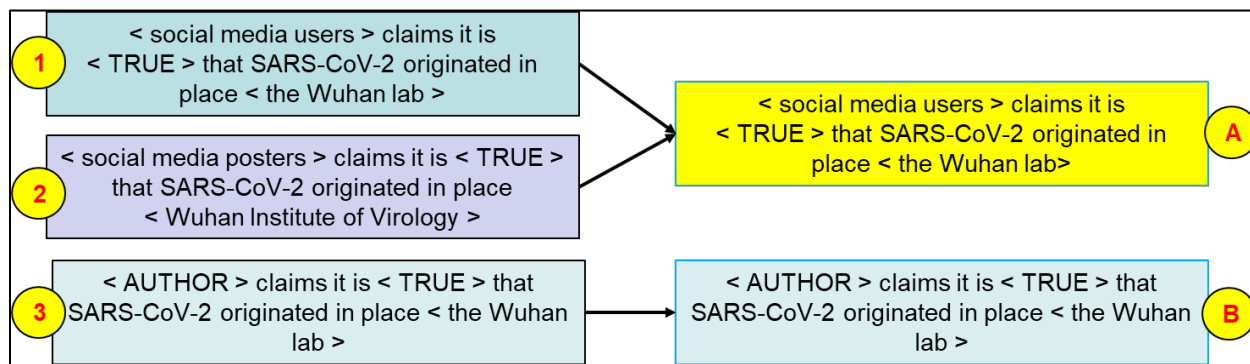


Figure 10: Clustering Duplicate Claims

When the de-duplicated claims are paired, each is assigned a label, either Claim A or Claim B. Relation decisions are made in terms of how Claim A relates to Claim B. There are four kinds of relations annotators can apply to a claim pair: identical, refuting, supporting, related. Annotators assess the relationships between claims from a reasonable, common-sense perspective.

Two claims are considered identical if they are on the same topic and subtopic, and their X Variables and Epistemic Statuses match. If only the Claimer varies between two claims, they are considered identical since the foundational meaning of the claim is not altered by the claims having been made by two different entities. Claims that have been clustered into the same equivalence class are automatically considered identical. The refuting, supporting, and related claim relationships require that the two claims in a pair not be considered identical. The following statement must be true of refuting claim pairs: "If Claim A is true, Claim B cannot be true." For two claims to be in a supporting relationship, the following statement must be true: "If Claim A is true, Claim B is more plausible." Finally, related claims must be on the same topic, and must not be identical, refuting, or supporting. Table 7 below shows examples of the cross-claim relations.

Table 7: Cross-Claim Relation Type Examples

Claim A	Claim B	Cross-Claim Relation
<CDC> claims it is <true> that <Children> can catch COVID-19	<FDA> claims it is <true> that <Children> can catch COVID-19	Identical
<Author> claims it is <true> that <The U.S. Army> created SARS-CoV-2	<Author> claims it is <true> that <the Chinese government> created SARS-CoV-2	Refuting
<Author> claims it is <true> that <Drinking lemon juice> prevents COVID-19	<Author> claims it is <true> that <Drinking acidic liquids> prevents COVID-19	Supporting
<Author> claims it is <true> that <Pets> transmits/transfers COVID-19	<Author> claims it is <true> that <Eating Chinese food> transmits/transfers COVID-19	Related

3.7 Quality Control

Quality control was performed on the annotation at several points in the process. First, the annotation interface prevents certain types of badly formed annotations from being created (missing required elements, combining incompatible types and subtypes, etc.). Then, after each stage of annotation, a second annotator reviews the annotations and corrects any errors. Once KB linking has been performed, additional corpus-wide quality checks are performed, such as checking all KB links for annotations with identical strings to ensure that they are linked consistently and reviewing the strings and types for all entities linked to the same KB node to check that there are no clear errors.

3.8 Assessment

Assessment was a key component of the AIDA data effort for all three phases. In Phase 1, TA1, TA2, and TA3 were all evaluated via assessment. In Phase 2 and Phase 3, TA1 was evaluated via automatic scoring against the gold standard human reference evaluation (see section 4.4 describing the approach to within-document annotation), while TA2 and TA3 were evaluated via assessment. The following sections describe TA1, TA2, and TA3 assessment procedures for each phase of AIDA. Information about how the assessments carried out by LDC were used in the scoring of system responses were provided to teams in advance of each evaluation (NIST, 2019, 2020, 2022).

3.8.1 TA1 Assessment

TA1 assessment for Phase 1 had two components: class assessment and graph assessment. The class assessment task asked assessors to judge whether a given system-extracted entity had the correct type assigned. Assessors were also required to complete coreference and KB linking of correct system responses to facilitate downstream tasks. During graph assessment, assessors

examined system-returned relation and/or event arguments and indicated whether the system's predicate justification provided adequate evidence for the assigned type, the argument role, and the entity filling the argument role. This task required close inspection of the multi-media provenance returned by systems to justify their events and relations. Graph assessment took place one argument role at a time, so assessors were not looking at a fully annotated event with all its roles, but instead were given a single event-role combination (e.g., Conflict.Attack Target) and its filler (e.g. MH17) and asked whether the justification supported the assertion of the filler having the given role in an event of the given type. Figure 11 shows an example of a justification and several correct argument roles and fillers that would be supported by that justification span and therefore would be marked correct in assessment.

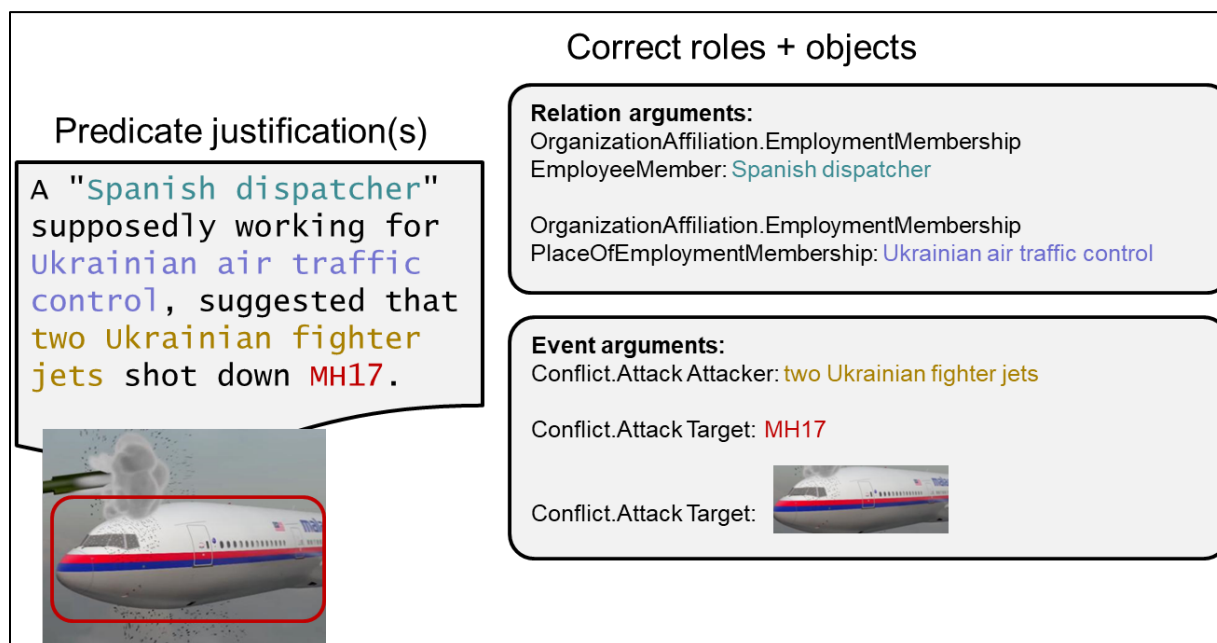


Figure 11: Graph Assessment Example

No assessment of TA1 system output was performed for Phase 2 and Phase 3 due to the shift to evaluation against gold standard reference annotation.

TA1 assessment results are discussed in section 5.7.

3.8.2 TA2 Assessment

In all three phases of AIDA, TA2 output was assessed by judging whether system-returned entities were correct responses to a query that identified the entity of interest. In Phase 1, TA2 output was evaluated via an assessment task referred to as "zero-hop" assessment. Assessors judged whether a given system entity was assigned by the system to the correct KB node (that is, was it conferral with the reference entity assigned to that KB node). In Phase 2 and Phase 3, namestring queries were added to the zero-hop task. For KB node queries, the task was identical to Phase 1. For namestring queries, assessors reviewed system-extracted entity mentions, and provided judgments on whether or not each mention contained a reference to an entity that matched the namestring in the query. Namestring queries may be correctly answered by mentions of multiple entities, unlike KB node queries in which any confusable names are disambiguated by the KB link. Therefore,

assessors also performed global clustering of all correct responses for namestring queries. In Phases 1 and 2, KB node queries used KB IDs from the LDC-produced reference KB; in Phase 3, KB node queries used Q node IDs from DWD, which served as the reference KB for Phase 3. Figure 12 shows several examples of responses for KB node and namestring queries.

		Is this a mention of “Anthony”?	Assessment	NIL Clustering
Is this a mention of Q148?	Assessment		yes	Cluster 1
China	yes		yes	
Wuhan	no		yes	Cluster 2
the entire country	yes	immunologist	yes	
		Anthony Bourdain	yes	

Figure 12: TA2 Zero-Hop KB Node (left) and Namestring (right) Queries

On the left side of the figure above, an example of a KB node query and assessment judgments are shown. The query uses a DWD Q node Q148 “People’s Republic of China”. Both the named mention “China” and the nominal mention “the entire country” (which document context shows is used to refer to China) are marked correct, while “Wuhan” is marked incorrect.

On the right side, there is an example of a namestring query. Rather than a KB node which distinguishes a specific real-world entity, only a namestring is provided. Namestrings may be full names, nicknames, or partial names (in this example, only a first name is used). For the namestring “Anthony”, and mention of someone named Anthony is considered correct, so all three of the responses in the example are correct. The correct answers are then grouped into two clusters, one referring to Anthony Fauci and one referring to Anthony Bourdain. Note that the two mentions of Fauci (one image and one nominal text mention) do not contain the string “Anthony”, but they are correct because they refer to a real-world entity named Anthony.

3.8.3 TA3 Assessment

The most complex assessment task, and the task most closely aligned with the overall AIDA program goals and use-case, was hypothesis assessment. In Phase 1 and Phase 2, this task consisted of judgements about the relevance of system events, relations and their arguments included in the hypothesis, relative to a Statement of Information Need that prompted the hypothesis as well as a reference perspective or prevailing theory. In Phase 3, this task evolved into assessment of claim frames and cross-claim relations.

In Phase 1, hypothesis assessment consisted of three parts: relevance, coherence, and coverage. Relevance assessment was a judgement about whether system events, relations and their arguments included in the hypothesis were relevant to a Statement of Information Need (SIN) which prompted the hypothesis. In other words, assessors judged whether the components of the hypothesis were responsive to the SIN. For example, given a SIN about attacks on protestors at Maidan Square, an attack event in which snipers shoot protestors at Maidan and a protest event in

which local residents participate in protests at Maidan Square would both be relevant, but a protest event in which protesters gather in Times Square would not. The Times Square protest event could be present in some document in the corpus, but it is not relevant to the attacks at Maidan Square, so would be marked as irrelevant. Next, assessors judged the semantic coherence of each hypothesis (i.e., is the hypothesis free of illogical or contradictory information). Either whole events or relations or individual arguments could be marked as incoherent. An example of a hypothesis that would have elements marked as incoherent would be one in which the hypothesis contains events in which the same entity is involved in two events at geographically separated locations at the same point in time, since the entity could not be in both places at once. Finally, assessors judged whether the system adequately covered the events and relations that underlie the prevailing theory associated with the hypothesis. This was done by comparing the full list of events/relations and arguments in the hypothesis with the list of events/relations and arguments in the prevailing theory and marking them as matched when they cover the same information.

In Phase 2, hypothesis assessment happened in two stages: hypothesis verification and knowledge element analysis. In hypothesis verification, assessors reviewed all system-produced hypotheses and identified any and all human-generated topic-specific perspectives that fully or partially matched each hypothesis. Figure 13 shows an example of three system hypotheses, two of which match a perspective, and one of which is unmatched.

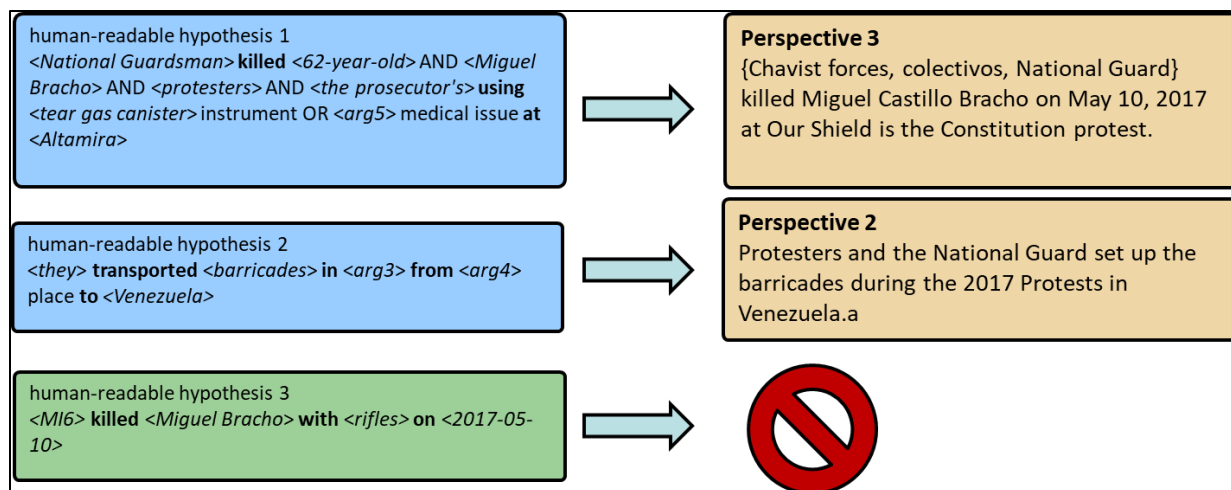


Figure 13: Matching Hypotheses to Perspectives

Any unmatched system hypotheses were reviewed for relevance and informativeness, as shown in Figure 14 below. The first hypothesis shown in the figure is relevant to the query about who killed Miguel Castillo Bracho, and it contains enough specific information to provide new, useful information not covered in one of the perspectives. The second hypothesis in Figure 14 is an example of a system hypothesis that is not relevant or informative, since the information it contains is so vague that it does not provide a meaningful answer to any query.

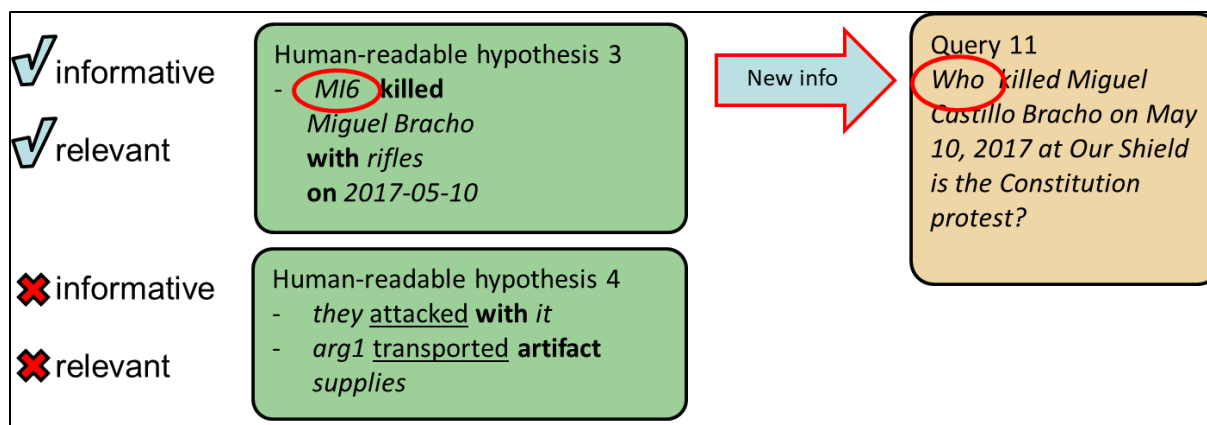


Figure 14: Evaluating Unmatched Hypotheses

Once the hypotheses have been matched with perspectives, the hypothesis moves to the next stage: knowledge element analysis. In this stage, assessors performed a detailed comparison of each hypothesis-perspective pair that was matched in hypothesis verification. For each pair, assessors determined a mapping between the constituent events, relations, and entities (i.e., knowledge elements) of the hypothesis and those of the perspective. They examined each individual knowledge element in the hypothesis and tried to find a matching knowledge element in the perspective. If no matching element was found, assessors determined whether the novel system knowledge element(s) provided accurate information that could be used to augment the human perspective. To judge accuracy of the novel elements, assessors examined the corresponding system-returned relation and/or event arguments and indicated whether the system's predicate justification provided adequate evidence for the assigned type, the argument role, and the entity filling the argument role. This task required close inspection of the multi-media provenance returned by systems to justify their events and relations.

The figures above provide simplified versions of the human readable hypotheses, but assessors had access to a version that contained additional details of each of the events and relations that made up the hypothesis. This human-readable hypothesis was the primary tool assessors used in comparing hypotheses to perspectives and allowed assessors to review an entire hypothesis in a practical, comprehensible "summary view" (see Figure 15).

```

event-129
Life.Die.DeathCausedByViolentEvents
Time T1 (start on or after): 2017-05-10
Time T2 (start on or before): 2017-05-10
Time T3 (end on or after): 2017-05-10
Time T4 (end on or before): 2017-05-10
<National Guardsman> killed <62-year-old> AND <Miguel Bracho> AND <protesters> AND <the prosecutor's>
using <tear gas canister> instrument OR <arg5> medical issue at <Altamira>
arg1 Killer National Guardsman PER.Combatant entity-83
arg2 Victim 62-year-old PER entity-23149
arg2 Victim Miguel Bracho PER entity-23155
arg2 Victim protesters PER.Protester entity-23156
arg2 Victim the prosecutor's PER.ProfessionalPosition entity-208
arg3 Instrument tear gas canister WEA entity-215
arg4 Place Altamira LOC entity-391
arg5 MedicalIssue n/a n/a n/a

```

Figure 15: Detailed Human-Readable Hypothesis

In Phase 3, hypotheses were represented as claim frames, with cross-claim relations between them. The three core TA3 assessment tasks were: assessment of the correctness of system claims and their constituent knowledge elements; clustering of (correct) system claims; and manual annotation of cross-claim relations between correct system claims and human-generated query claims. The latter task was similar to the cross-claim relation annotation task. No explicit alignment of system claims to pre-existing reference perspectives was performed (in contrast to prior phases). Figure 16 below shows the Phase 3 TA3 assessment pipeline.

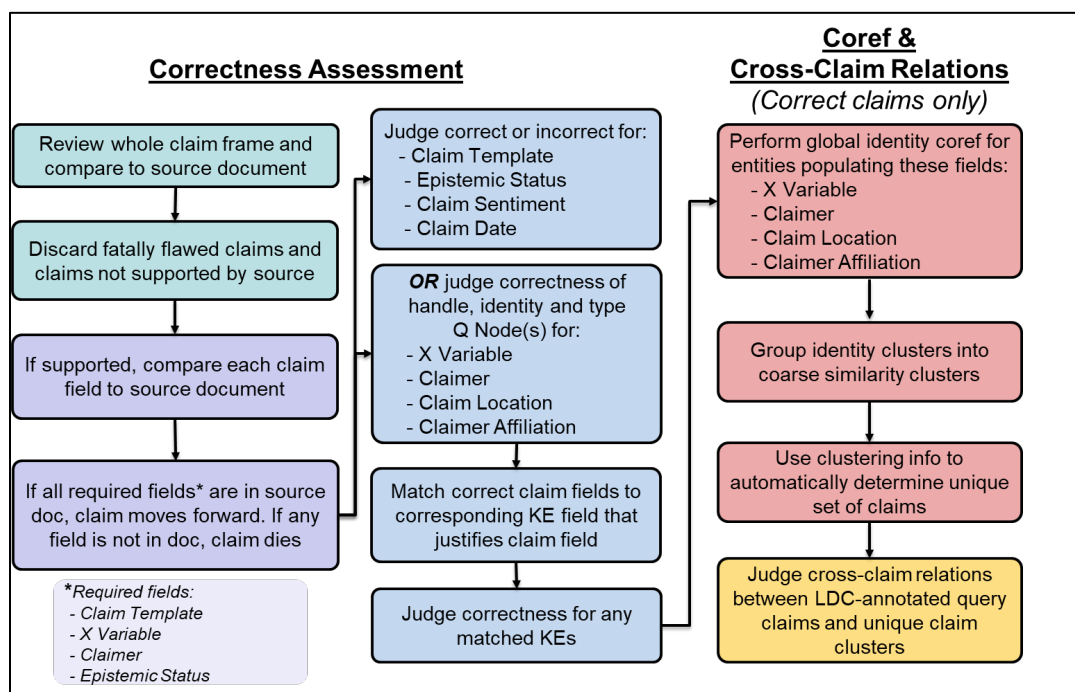


Figure 16: Full Phase 3 Assessment Pipeline for TA3

In the first stage of TA3 assessment a system-produced claim frame is judged against a source document to determine whether the claim is correct with respect to the information in that document. Consider the example claim below, and the text extract returned by a system as justification for the claim:

- **Claim:** <Cat Antonio> claims it is TRUE that <the French> created SARS-CoV-2
- **Document text:** *A new conspiratorial video massively affirms that the coronavirus was created “in laboratory” and “patented.” In the video, originally posted on Facebook by the account of a certain “Cat Antonio,” the man explains that the government is lying about the Covid-19 epidemic, that the virus was “patented” and “created by the French and the Pasteur Institute.” This is not the case and the video which disseminates this message is based on rumors and the misunderstanding of a document.*

The first step in correctness assessment is to judge whether the claim makes enough sense to be further assessed (it does), and whether the claim, as a whole, is at least partially justified in the document (it is). Each aspect of the claim is then judged separately, proceeding through a priority-ranked list of the claim fields. Subsequent claim fields do not get assessed if higher priority claim

fields were marked incorrect. For instance, the X Variable is the top-ranked field in the list, and so if that element of the claim is not supported by the document, there is no need to assess lower priority fields (such as Claimer, Claim Location, etc.). In the example above, the document supports the claim that “the French” created SARS-CoV-2, so the X Variable is correct and additional fields are assessed. The reason for the ranked list of fields is that basic information such as the X variable, epistemic status, and claimer are so central to the meaning of a claim that it does not make sense to ask whether things like claim location or claimer affiliation are correct in cases where the basic content of the claim is not justified in the document. Claims that were considered correct overall in the first stage of TA3 assessment move onto the second and third stages. A claim is considered correct if all four of the highest priority fields (Claim Template, X Variable, Claimer, Epistemic Status) are assessed as correct.

The second stage of TA3 assessment involves cross-document clustering of the entities populating claim fields. The first step of clustering is done through automatic or manual global identity coreference for: X Variable, Claimer, Claim Location, Claim Medium, and Claimer Affiliation. For fields with correct KB links (which is known from the first stage of assessment), identity coreference can be done automatically. For fields without correct KB links (either because the system assigned an incorrect link or because the KB does not contain a node for that entity), identity coreference must be done manually by assessors.

After the strict identity coreference step, a second clustering pass is performed that involves grouping strict identity clusters into coarse equivalence classes. Assessors can place 2 or more identity clusters into the same coarse equivalence class if the meanings of the clusters are roughly equivalent. For example, an identity cluster containing mentions of “lemon” and an identity cluster containing mentions of “lemon juice” might be grouped in the same coarse equivalence class.

Figure 17 below shows examples of identity coreference and coarse-grained clustering in TA3 assessment.

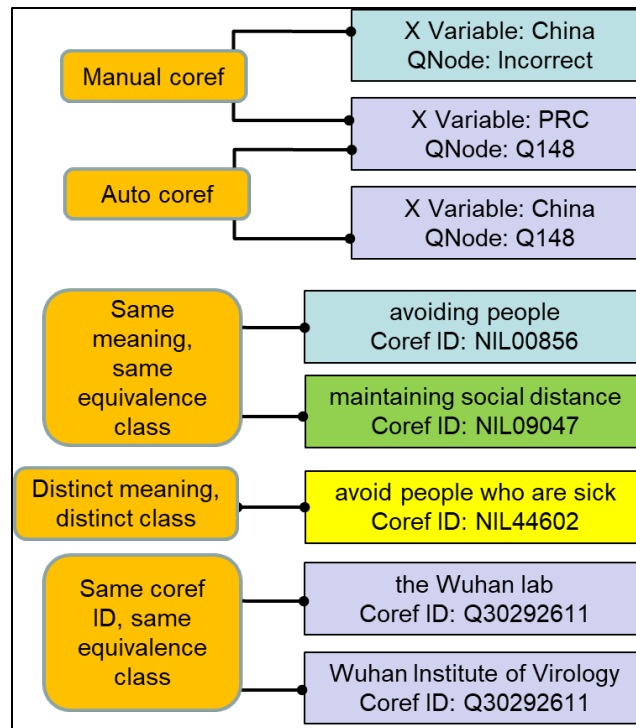


Figure 17: Identity Coreference and Coarse-Grained Clustering for Claims

The third and final stage of TA3 assessment involves the manual creation of cross-claim relations between correct system claims and human-annotated query claims. In order to reduce redundancy and ensure consistent handling of individual claims, cross-claim relations are done for claim clusters using the coreference information determined in the previous assessment stage. Equivalence classes are created for system claims using Claim Template, X Variable, and Epistemic Status polarity (true/false). For each system claim-query claim pair, assessors determine if the system claim supports or refutes the query claim.

3.9 Support for Similarity-Based Scoring

One of the main challenges associated with the transition from a constrained tagset to the open DWD ontology was the need to find a way to automatically score the semantic similarity of entities/relations/events/claims tagged with different Q nodes, since there could be multiple reasonable ways to represent the same information. LDC participated in this program goal in two ways: providing manual similarity judgements of DWD concepts to support development of similarity metrics by AIDA teams working with the Ontology Working Group, and developing an approach to using the similarity metrics in IAA scoring on annotations.

To support the development/refinement of similarity measures by ADIA teams, LDC collected lenient and strict annotator judgments for pairs of DWD concepts by trained LDC annotators, as well as entirely untrained crowdworker judgments. Lenient judgments by LDC annotators followed the AIDA assessment standard that had been applied in previous phases, where types were not required to match exactly if the same information was conveyed. Strict judgments required annotators to use very literal and narrow reading of the definition of each concept. Crowdworkers

were given two examples to calibrate their understanding of the required judgments, but no other training.

A total of 250 concept pairs were judged, 50 pairs taken from previous AIDA evaluation data where systems and annotators had applied different tags to the same string, and 200 additional thematic pairs selected from DWD (e.g., profession/role-related pairs). All pairs were judged in the abstract (label + definition), not in the context of mentions grounded in data. Figure 18 shows an example of the judgement task.

How close in meaning are these two items?
Instructions:
Review the two entities and their definitions and indicate how close they are in meaning.
Example 1

1) per.politician.headofgovernment: A head of government, a person who rules or commands
2) per.professionalposition.minister: A government minister, a person appointed to a high office in the government
5: Very similar
4: Similar
3: Slightly similar
2: Dissimilar
1: Totally dissimilar and unrelated

Figure 18: Similarity Judgment Task for Tagset Concepts

In addition to supporting the general program goal of similarity-based scoring, LDC also needed to update our Inter-annotator agreement (IAA) scoring tools to utilize this approach. The first step was to update existing LDC IAA software to utilize the entire DWD tagset, and then to use the Ontology Working Group mapping of LDC tags to DWD Q nodes to convert LDC tags to Q nodes and score using both an exact match and a similarity-informed match standard. In order to understand the impact of a fully open ontology, we also produced a set of entity annotations for a very small set of documents in which annotators were allowed to use any tag from DWD, not just those mapped to the LDC tagset. Results of this experiment are presented in section 5.8.

The final task involved in supporting similarity scoring was also the most challenging. Moving away from comparing mentions anchored in specific text spans toward being able to score entire claim frames, including all their component slots, to gauge agreement using a similarity-based measure. To calibrate how well this worked, LDC annotators manually judged 1,990 pairs of claims for similarity. These manual comparisons were then compared to early output from the IAA scorer. Qualitative investigation of cases in which humans gave a pair of claims high similarity but the software did not map them at all showed that many were caused by the tool's overly strict requirements for template matching for topics that were closely related. For example, a human assessor might give the following two claims high similarity ratings:

- < Fort Detrick laboratory> created SARS-CoV-2
- SARS-CoV-2 originated in < US Army Medical Research Institute of Infectious Diseases at Ft. Detrick>

But because one has the template “X created SARS-CoV-2” and the other uses “SARS-CoV-2 originated in place-X”, the software initially disallowed these from being mapped. Loosening the requirements for template matching improved the consistency of the automatic IAA scoring with the human judgments.

The scoring tool that resulted from the work on similarity-based scoring was used to calculate the Phase 3 agreement scores included in the annotation results sections (section 5.4-5.6). The tool assigns each pair of entity, event, or relation mentions, or each pair of claims, a score [0,1]. For entity, relation, and event mentions grounded in document spans, a minimum threshold of overlap (0.1) in the span prevents completely unrelated mentions from being mapped to each other. Overall mapping is done using Hungarian algorithm for minimum weight matching in bipartite graphs (SciPy, 2016). The mapping and scoring of entity mentions is done first, so that event and relation mapping can take advantage of argument match information. Score calculation for events and relations uses a weighted average of span overlap, qnode_type similarity, and entity mapped scores for arguments. For claims, which do not have spans that can be used to guide mapping, a weighted average of claim template, epistemic status, X-variable similarity (using qnode_id, qnode_type, and the text description).

4 Results & Discussion

4.1 Scenario and Topic Development

4.1.1 Scenario and Topic Selection

Each of the subsections below presents the results of scenario and topic selection in the form of details about the selected topics and designation of practice and eval topics for each phase of AIDA.

4.1.1.1 Phase 1: Russia-Ukraine Relations

The Russia-Ukraine Relations scenario that was the focus of AIDA Phase 1 centered on political and military relations between Russia and Ukraine in the 2014-2015 period, including events leading up to and following the main focal topics. The scenario was chosen because it was rich in opposing narratives and portrayals of the facts, across a variety of media types and languages. Within the broader domain, several subtopics were selected for data collection and annotation.

An initial set of nine scenario-relevant topics were selected for Phase 1 of AIDA and used in the first part of Phase 1, leading up to the Month 9 Pilot Evaluation; after the Month 9 evaluation, a subset of six topics were selected for re-annotation and use in the Month 18 Phase Evaluation. The table below shows the names of each topic, as well as the division of topics in to practice topics for use in system development and eval topics, which were not revealed until the start of the evaluation.

Table 8: Phase 1 Topics

Set	Topic
Practice	Crash of Malaysian Air Flight MH17 (Pilot Only)
Practice	Flight of Deposed Ukrainian President Viktor Yanukovich (Pilot Only)
Practice	Humanitarian Crisis in Eastern Ukraine (July-August 2014) (Pilot Only)
Practice	Who Started the Shooting at Maidan? (February 2014)
Practice	Ukrainian War Ceasefire Violations in Battle of Debaltseve (January-February 2015)
Practice	Donetsk and Luhansk Referendum, aka Donbas Status Referendum (May 2014)
Eval	Suspicious Deaths and Murders in Ukraine (January-April 2015)
Eval	Odessa Tragedy (May 2, 2014)
Eval	Siege of Sloviansk and Battle of Kramatorsk (April-July 2014)

4.1.1.2 Phase 2: Crisis in Venezuela

The scenario selected for AIDA Phase 2 was Crisis in Venezuela, which covered the political, economic, and social upheavals that occurred in Venezuela after the death of Hugo Chávez in 2013 and during the early years of the presidency of Nicolás Maduro through to 2017. Like the Phase 1 scenario, it was selected for its wealth of informational conflict across media types and languages. Within the broader domain, six subtopics were selected for data collection and annotation, three as practice data for systems and three as data for evaluation. LDC and MITRE worked together to partition topics so new domain areas and ontology concepts needed in eval topics were likely to occur in practice topics. The partition is shown in Table 9 below.

Table 9: Phase 2 Topics

Set	Topic
Practice	2014 Disease Outbreak in Venezuela
Practice	2017 Venezuelan Constituent Assembly Election
Practice	Drone Explosions in Caracas
Eval	Death of Hugo Chávez
Eval	2017 Protests in Venezuela
Eval	Helicopter Attack in Caracas

4.1.1.3 Phase 3: COVID-19 Global Pandemic

The AIDA Phase 3 scenario, COVID-19 Global Pandemic, covers many domains of misinformation, disinformation, and variable truth in the context of the burgeoning COVID-19 pandemic in 2020 and early 2021. Its topics focus on the inception and spread of the SARS-CoV-2 virus, the nature of the disease, and the effectiveness or consequences of treatments for the disease. The scenario contains a wealth of informational conflict across media types with an increased focus on the variation of truth over time as information evolves compared to previous scenarios. Additionally, the Phase 3 scenario allows for exploration of parallel, non-conflicting answers to queries that could be mistaken for informational conflict in some contexts. Within the scenario, several broad topics were selected, and each of these topics was further divided into subtopics for data collection and annotation.

The list of topics and subtopics for Phase 3 is substantially larger than the lists for the previous two phases, reflecting the new emphasis on claim frames, which requires a focus on a single point of informational conflict per claim. Each subtopic is also associated with a claim template that uses an “X” to describe the variable that is expected to contain informational conflict for that subtopic during Claim Frame Annotation. The breakdown of topics, subtopics, claim templates, and their partitions are described in Table 10 below.

Table 10: Phase 3 Topics

Set	Topic	Subtopic	Claim Template
Practice	Characterizations of the Virus	Characterizations of coronavirus	SARS-CoV-2 is X
Practice		Immunity to the virus	Immunity against COVID-19 is X
Practice		Virus is losing its potency	The potency of SARS-CoV-2 is X
Eval	Contracting the virus	Decreased risk	X will decrease the risk of contracting COVID-19; X will NOT decrease the risk of contracting COVID-19
Eval		Increased risk	X will increase the risk of contracting COVID-19; X will NOT increase the risk of contracting COVID-19
Practice		Who contracts the virus	X can catch COVID-19
Eval	COVID-19 vaccine is harmful	COVID-19 vaccine causes medical conditions	COVID-19 vaccine causes X medical condition
Eval		COVID-19 vaccine is causing people to die	X died because of the COVID-19 vaccine

Eval		A SARS-CoV-2 variant is associated with a location	Location/GPE-X has a SARS-CoV-2 variant
Eval	COVID-19 Variants	A SARS-CoV-2 variant is more dangerous than other strains	Variant-X of SARS-CoV-2 is more dangerous/lethal/virulent/transmissible than other strains
Practice	Curing/Preventing /Destroying the Virus	Curing the virus	X cures COVID-19
Practice		Destroying the virus	X destroys COVID-19
Practice		Preventing the virus	X prevents COVID-19
Eval	Detecting the Virus / Testing for the Virus	Accuracy of Tests	test X is accurate (enough) vs. test X is not accurate (enough)
Eval		Detecting the Virus	X will detect the virus vs. X will not detect the virus
Practice	Government actions related to the virus	Contact tracing related to the virus	Government-X enacted contact tracing related to COVID-19
Eval		Federal government seizing/diverting PPE and other medical supplies	Federal government is seizing/diverting PPE and other medical supplies from X
Practice		Government Cover-Ups about Cures	Government-X is withholding cures for COVID-19
Practice		Patient Containment actions related to the virus	Government-X enacted patient containment actions related to COVID-19
Practice		Population restrictions related to the virus	Government-X enacted population restrictions related to COVID-19
Eval	Non- Pharmaceutical Interventions (NPIs)	NPI Effectiveness against COVID-19	NPI X is effective against COVID-19
Eval		NPI Harmfulness	NPI X is harmful; NPI X is NOT harmful
Eval		NPI Necessity for preventing the spread of COVID-19	NPI X is necessary for preventing the spread of COVID-19

Eval		NPI Recommendations for preventing the spread of COVID-19	NPI X is recommended for preventing the spread of COVID-19
Eval		NPI Regulations for preventing the spread of COVID-19	NPI X is required for preventing the spread of COVID-19
Practice	Non-Pharmaceutical Interventions (NPIs): Masks	Mask Harmfulness	Wearing masks has X negative effect
Practice		Mask Necessity	Wearing masks is necessary in location/GPE/facility-X
Practice		Mask Recommendations	Wearing masks is recommended in location/GPE/facility-X
Practice		Mask Regulations	Wearing masks is required in location/GPE/facility-X
Practice	Origin of the Virus	Animals associated with the origin of the virus	Animal-X is associated/involved with the origin of COVID-19
Eval		Events associated with the origin of the virus	Event-X is associated/involved with the origin of COVID-19
Eval		What caused the virus outbreak	The cause of the origin of COVID-19 is X
Practice		When the outbreak originated	The first case of COVID-19 occurred on date-X
Practice		Where the virus first occurred	The first case of COVID-19 occurred in location-X
Practice		Where the virus originated	SARS-CoV-2 originated in place-X
Eval		Who was/were the first patient(s)/victim(s)/carrier(s) of the COVID-19 virus?	Person-X was the COVID-19 index case
Practice		Who/what was the target of the virus	X was the target of the virus
Practice		When the virus was created	SARS-CoV-2 was created on date-X

Practice	Origin of the Virus: Virus Creation	Where the virus was engineered/created	SARS-CoV-2 was created in location/GPE/org/facility-X
Practice		Who created the virus	X created SARS-CoV-2
Practice		Who funded the creation of the virus	X funded SARS-CoV-2 development
Eval	Transmitting the virus	How virus is transmitted	COVID-19 can be transmitted in X conditions; COVID-19 can NOT be transmitted in X conditions
Practice		What transmits the virus	X transmits / transfers COVID-19
Practice	Treatment availability	Treatments approved for Emergency Use Authorization	Treatment-X received emergency use approval
Practice		Effective treatments for COVID-19	X is an effective treatment for COVID-19
Practice		Treatments that prevent or reduce the chance of death from COVID-19 infection	Treatment-X prevents death or reduces the chance of death from COVID-19 infection
Practice		Treatments that shorten the length of COVID-19 infection	Treatment-X shortens the length of COVID-19 infection
Practice	Treatment safety	Generally safe medications that are unsafe for COVID-19 patients	Generally safe medication X is unsafe for COVID-19 patients
Practice		Safe treatments for COVID-19	Treatment-X is a safe treatment for COVID-19
Eval	Using COVID-19 to implant tracking microchips	Using COVID-19 to implant tracking microchips	X is planning to implant tracking microchips via COVID-19 vaccines; X is NOT planning to implant tracking microchips via COVID-19 vaccines
Eval	Who is Sick / Who has tested positive	Who has/had COVID-19	X has/had COVID-19
Eval	Who profits or benefits from the disease	Who profits or benefits from the disease	X benefits from the disease because of Y; X does not benefit from the disease because W

4.2 Data Collection

In Phase 1, data was collected in English, Russian, and Ukrainian. The collected data was partitioned into three parts: a background corpus of English, Russian, and Ukrainian data from the approximate time period of the topics of interest; a corpus of documents collected specifically for their relevance to the topics designated as practice topics for use in system development, and an evaluation corpus consisting of some documents specifically collected for their relevance to the topics designated as evaluation topics, as well as background data that is from the same approximate time period and similar sources, some of which may also have general relevance to the wider scenario of Russia-Ukraine relations. The table below shows the number of documents in each partition.

Table 11: Phase 1 Document Counts by Language

	Eng	Rus	Ukr	Mixed	Total
Background	783	1,753	1,101	6,031	9,668
Practice	245	518	269	639	1,671
Evaluation	752	511	670	63	1,996
Total	1,780	2,782	2,040	6,733	13,335

Table 12 shows the number of individual text, image, and video files present in each partition. Each collected document typically has at most one text element, but may have multiple image and/or video elements, such that the number of files exceeds the number of documents in the corpus.

Table 12: Phase 1 File Counts by Media Type

	Text	Image	Video	Total
Background	8,189	13,770	943	22,902
Practice	1,608	10,537	479	12,624
Evaluation	1,996	6,194	322	8,512
Total	11,793	30,501	1,744	44,038

Table 13 shows the count of documents by topic.

Table 13: Phase 1 Document Counts by Topic

Topic	Documents
Who Started the Shooting at Maidan? (February 2014)	128
Ukrainian War Ceasefire Violations in Battle of Debaltseve (January-February 2015)	104
Donetsk and Luhansk Referendum, aka Donbas Status Referendum (May 2014)	100
Suspicious Deaths and Murders in Ukraine (January-April 2015)	121
Odessa Tragedy (May 2, 2014)	128
Siege of Sloviansk and Battle of Kramatorsk (April-July 2014)	179

After Phase 1, large background collection was not required, so the data sets focused primarily on topic-relevant and scenario-relevant documents, with a small amount of background data included in evaluation corpora.

For Phase 2, with the shift of scenario from Russia-Ukraine Relations to the Crisis in Venezuela, Ukrainian was dropped as a target language and Spanish was added. English and Russian were both collected in Phase 2 as in Phase 1. The collected data was partitioned into two parts: a corpus of documents collected specifically for their relevance to the topics designated as practice topics for use in system development; and an evaluation corpus consisting of some documents specifically collected for their relevance to the topics designated as evaluation topics, as well as some generally scenario-relevant and background documents. The tables below show, respectively: the number of documents in each partition; and the number of documents per topic.

Table 14: Phase 2 Document Counts by Language

	Eng	Rus	Spa	Mixed	Total
Practice	465	343	511	194	1,513
Evaluation	425	506	814	311	2,056
Total	890	849	1325	505	3,569

The table below shows the number of individual text, image, and video files present in each partition. Each collected document typically has at most one text element, but may have multiple

image and/or video elements, such that the number of files exceeds the number of documents in the corpus.

Table 15: Phase 2 File Counts by Media Type

	Text	Image	Video	Total
Practice	1,340	8,674	337	10,351
Evaluation	1,806	6,448	535	8,789
Total	3,145	15,122	872	19,139

The table below shows the number of documents collected per topic.

Table 16: Phase 2 Document Counts by Topic

Topic	Documents
2014 Disease Outbreak in Venezuela	315
2017 Venezuelan Constituent Assembly Election	596
Drone Explosions in Caracas	589
Death of Hugo Chávez	211
2017 Protests in Venezuela	324
Helicopter Attack in Caracas	236

Finally, in Phase 3, the languages remained the same as the domain shifted to the COVID-19 Pandemic. With a larger number of less-complex topics suitable for claim frame annotation, the number of documents per topic was necessarily smaller than in previous phases, which had focused on fewer, more complex topics. Additionally, the decision was made at the program level to exclude video data from collection and annotation in Phase 3. Therefore, counts of files by media type for Phase 3 include only text and image.

The Phase 3 data was partitioned into two parts: a corpus of documents collected specifically for their relevance to the topics designated as practice topics for use in system development; and an evaluation corpus consisting of some documents specifically collected for their relevance to the topics designated as evaluation topics. The table below shows the number of documents in each partition.

Table 17: Phase 3 Document Counts by Topic

	Eng	Rus	Spa	Mixed / Unknown	Total
Practice	483	248	395	293	1,419
Evaluation	1,199	391	409	12	2,011
Total	1,682	639	804	305	3,430

The table below shows the number of individual text and image files present in each partition. Each collected document typically has at most one text element, but may have multiple image elements, such that the number of files exceeds the number of documents in the corpus.

Table 18: Phase 3 File Counts by Media Type

	Text	Image	Total
Practice	1,419	11,815	13,234
Evaluation	2,011	13,002	15,013
Total	3,430	24,817	28,247

The final table in this section shows the number of documents collected per topic for Phase 3.

Table 19: Phase 3 Document Counts by Topic

Topic ID	Topic Name	Count
C301	Contracting the virus – increased risk	35
C302	Contracting the virus - decreased risk	42
C303	Who contracts the virus	67
C304	What transmits the virus	64
C305	How virus is transmitted	31
C306	Destroying the virus	65
C307	Preventing the virus	63
C308	Curing the virus	77
C309	Detecting the Virus	48
C310	Accuracy of Tests	38
C311	Characterizations of coronavirus	53
C312	Immunity to the virus	53
C313	Virus is losing its potency	62
C314	NPI Regulations	53
C315	NPI Recommendations	65
C316	NPI Effectiveness	68
C317	NPI Necessity	69
C318	NPI Harmfulness	58
C319	Animals associated with the origin of the virus	66
C320	Who was/were the first patient(s)/victim(s)/carrier(s) of the COVID-19 virus?	69
C321	Events associated with the origin of the virus	47

C322	Who created the virus	86
C323	Who funded the creation of the virus	52
C324	Where the virus originated	58
C325	Where the virus was engineered/created	68
C326	Where the virus first occurred	49
C327	When the outbreak originated	72
C328	When the virus was created	49
C329	Who/what was the target of the virus	35
C330	What caused the virus outbreak	51
C331	Population restrictions related to the virus	46
C332	Contact tracing related to the virus	48
C333	Patient Containment actions related to the virus	50
C334	Federal government seizing/diverting PPE	36
C335	Government Cover-Ups about Cures	35
C336	Who is sick / Who has tested positive	50
C337	Who profits or benefits from the disease	43
C338	Using COVID-19 to implant tracking microchips	45
C343	Effective treatments for COVID-19	59
C344	Treatments that shorten the length of COVID-19 infection	28
C345	Treatments that reduce the severity of COVID-19 infection	19
C346	Treatments that prevent or reduce the chance of death from COVID-19 infection	48
C347	Treatments being withheld from the public	7
C348	Treatments approved for Emergency Use Authorization	11

C349	Safe treatments for COVID-19	34
C350	Generally safe medications that are unsafe for COVID-19 patients	19
C358	COVID-19 vaccine is causing people to die	27
C359	COVID-19 vaccine causes medical conditions	48
C360	Dangerous variants	8
C361	COVID-19 variant is associated with a location	7
Total	-	2,381

4.3 Ontology/Tagset Development

With the interaction between the expansion of the tagset for annotation with scoping of annotation to accommodate different degrees of exhaustive mention annotation for TA1, it is interesting to compare the number of unique types tagged in each phase.

The table below shows the total number of unique entity, relation and event types for each phase (defined as a unique type.subtype.subsubtype combination for Phases 1 and 2 and a unique DWD node for phase 3). For Phase 3, there is an additional category of “claims” for the various slots in the claim frame template that were filled by Q nodes, since these were not categorized as entities, relations, or events.

Table 20: Unique Tags Annotated in by Phase

Phase	Entity	Relation	Event	Claims	Total
Phase 1 Pilot	14	16	21	n/a	51
Phase 1 Final	163	47	116	n/a	326
Phase 2	110	38	75	n/a	223
Phase 3	2169	46	76	1025	3208

The large increase in unique types annotated between the pilot and final Phase 1 is due to the increased size of the tagset. Of note, in spite of the move to more exhaustive annotation in Phase 2, the total number of unique types tagged was slightly smaller, most likely due to the scoping by document region that was necessary to make the more exhaustive annotation feasible. In phase 3, the number of unique tags increased tenfold. This was due to both the expansion to the use of fine-grained type tags from DWD for entities as well as the large number of slots in claim templates that could be filled by Q nodes. This progression shows that annotation from an extremely large

ontology such as DWD can result in a massive variety of tag types being used in annotation, and such an approach is feasible given realistic expectations about exhaustiveness.

4.4 Within-document Annotation

The total amount of annotation completed in each phase is summarized in the table below, which shows the number of documents annotated, as well as the number of entity, event, and relation mentions labeled in each data set. For Phase 3, the number of claim frames annotated is also included.

Table 21: Annotated Documents and Mentions by Phase and Practice/Eval Partition

	docs	entity	relation	event	claims
Phase 1 Pilot	704	4,529	569	1,967	n/a
Phase 1 practice topics	204	8,930	1,848	1,955	n/a
Phase 1 eval topics	248	10,508	2,029	2,456	n/a
Phase 2 practice topics	30	3,374	360	560	n/a
Phase 2 eval topics	72	4,382	407	1,061	n/a
Phase 3 practice topics	65	13,380	711	2,112	638
Phase 3 eval topics	53	7,651	645	1,798	440
Total	1,376	52,754	6,569	11,909	1,078

In the Phase 1 and Phase 2 IAA was reasonably high given the relatively large size of the tagset. We scored obtained pairwise scores for annotators and achieved a maximum f-score of 94.67 for entities, 86.84 for events, and 92.74 for relations. The average f-score for all pairs of annotators scored was 82.98 for entities, 78.00 for events, and 64.63 for relations.

In Phase 3, we moved to similarity-based scoring, and obtained IAA scores using the scoring tools discussed in section 4.9. Table 22 shows the scores when mapping of mentions is constrained by different levels of similarity. When mentions are allowed to map to each other with a “good” similarity score (above 0.5), scores are a bit higher than when they are allowed to map to each other only when they gave a “very good” similarity score (over 0.75). The table includes results for text and image together on the top row, text only on the second row, and image only on the third row.

Table 22: Similarity-Based IAA Scores for Entities, Relations, and Events

	Entities		Events		Relations	
	>=0.5	>=0.75	>=0.5	>=0.75	>=0.5	>=0.75
text+box	89.93	82.87	63.86	55.42	62.07	62.07
text only	93.48	92.66	67.19	57.81	62.07	62.07
box only	78.41	51.10	52.63	42.11		

Relations are rare in image data generally, so the lack of a score for relations on the third row reflects the lack of annotations of relations in the images that were scored for IAA. In general, however, image annotations score worse than text, which is not surprising since images can have a lot of edge cases (e.g., small or partially obscured objects, or objects seen at a distance). In the version of the scoring presented here, text spans are heavily weighted for entities, so scores would be expected to improve with more forgiving handling of differences such as “the CIA” vs “CIA”.

4.5 Cross-document Annotation

The following table describes the number of KB links and NIL mentions linked to the reference KB or assigned to NIL clusters during coreference in both practice and evaluation partitions across all three phases. As part of the trade-off between providing exhaustive annotation for TA1 and maintaining an appropriate amount of resources available to devote to TA3, cross-document coreference was not performed in Phase 2, although within-document NIL coreference clusters were provided.

Table 23: KB Links and NIL Coreference Decisions by Phase

	Practice Topics		Evaluation Topics		Total
	KB links	NILs	KB links	NILs	
Phase 1	4,678	4,249	5,700	4,805	19,432
Phase 2	-	3,373	-	4,381	7,754
Phase 3	2,930	6,815	2,281	6,423	18,449

4.6 Hypotheses

Direct annotation of hypotheses in data was limited to the Phase 1 Pilot stage and the Phase 3 Claim Frame annotation. In the final Phase 1 evaluation, and in Phase 2, hypotheses were represented as prevailing theories and perspectives, respectively, which did not require any annotations grounded in specific documents.

In Phase 1, each event or relation mention annotated was also judged for its relevance to the facet-level hypotheses described in section 4.6.1. This resulted in 45,240 judgments indicating whether events and relations supported or contradicted facet-level hypotheses.

In Phase 3, hypotheses were represented as claim frames, which were annotated directly. The following table shows the number of documents annotated for claim frames across the three languages for practice and evaluation topics.

Table 24: Annotated TA3 Documents by Language

	Total	English	Spanish	Russian
Annotated Practice Topic Documents	58	30	13	15
Annotated Evaluation Topic Documents	50	20	15	15

The annotation of claim frames required both the population of the slots in the claim frame template (such as X Variable, Claimer, Epistemic Status, etc.), but also the annotation of a set of events and relations, with their arguments, that could be used to support the claim frame. The number of claim frames, entities, relations, and events annotated in practice and evaluation topics for Phase 3 is shown in the table below.

Table 25: Annotation Counts for Claims, Entities, Relations, and Events in TA3

	Claim Frames	Entities	Relations	Events
Practice Topic Annotations	637	1150	413	930
Evaluation Topic Annotations	439	1054	357	735

IAA for claim frame annotation was measured using the method described in section 4.9. Since AIDA teams and LDC annotators independently annotated a common pool of documents, we compared agreement both within LDC's team and across teams. Within LDC, agreement ranged from a low of 62.16 to a high of 81.25 between different pairs of annotators when similarity at the "good" level was required for mapping. Cross-team IAA ranged from a low of 45.88 to a high of 51.78 with the same level of similarity requirement. Given that the teams were free to take very different approaches to annotation, sharing only a set of guiding principles and format definitions, the fact that LDC annotators agreed more closely with each other than any cross-team pairings should not be surprising.

The final stage of claim frame annotation was the application of cross-claim relation labels to pairs of claims. LDC annotated 10,320 cross-claim relations for practice topic claims and 29,056 cross-claim relations for evaluation topic claims, totaling 39,376 cross-claim relations for both partitions. The table below shows the breakdown of claim relations in the four categories of identical, refuting, supporting, and related.

Table 26: Count of Cross-Claim Relation Annotations by Type

	Identical	Refuting	Supporting	Related
Practice Topics	140	2,293	1,581	6,306
Evaluation Topics	29	506	2,728	25,793
Total	169	2,799	4,309	32,099

4.7 Assessment

Table 27 below shows the number of system responses assessed for each of the assessment tasks in Phase 1. In all assessment judgements, assessors followed a lenient assessment standard, giving credit for system responses that were not perfect (e.g., justification span is inexact but contains an appropriate mention). The class and graph assessments were performed on TA1 system responses, zero-hop assessment addresses TA2 system responses, and hypothesis assessment was done on TA3 system responses.

Table 27: Phase 1 System Responses Assessed

Assessment Task	System Responses Assessed
Class	8,859
Zero-hop	5,978
Graph	16,854
Hypotheses	9,949

Table 28 below shows the number of system responses assessed for each of the assessment tasks in Phase 2. In all assessment judgements, assessors followed a lenient assessment standard, giving credit for system responses that were not perfect (e.g., justification span is inexact but contains an appropriate mention).

Table 28: Phase 2 System Responses Assessed

Assessment Task	System Responses Assessed
Zero-hop	2,099
Hypothesis	47,843

Table 29 below shows the number of decisions made by LDC annotators for TA3 annotation and assessment in AIDA Phase 3.

Table 29: Phase 3 System Responses Assessed

Task	Decisions
TA2 Assessment	4,585
Hypothesis & claim frame annotation	34,953
TA3 assessment	13,506

Comparing the numbers of decisions related to hypotheses and TA3 assessment in all three phases makes clear that LDC's effort toward evaluation of AIDA systems during assessment shifted steadily to direct our effort toward the TA3 Hypothesis tasks and away from the diagnostic measures of TA1

and TA2 over the course of AIDA. In Phase 1, the majority of the assessment decisions were made on TA1 (class, graph) and TA2 (zero-hop) tasks, with less than a quarter of the effort directly devoted to hypothesis assessment. In Phases 2 and 3, over 90% of the effort went toward hypotheses directly.

4.8 Support for Similarity-Based Scoring

Results of the similarity judgement task for concepts were shared with the groups of AIDA performers working on similarity scoring. An example of the results is shown in Figure 19.

ontology pair		LDC Strict Judges					DC Lenient Judges					Crowdworkers				
Item 1	Item 2	5	4	3	2	1	5	4	3	2	1	5	4	3	2	1
per.politician.headofgovernment: A head of government, a person who rules or commands	per.professionalposition.spokesperson: A spokesperson	0	0	2	0	0	0	1	1	0	0	0	2	2	5	1
per.politician.headofgovernment: A head of government, a person who rules or commands	per.professionalposition.minister: A government minister, a person appointed to a high office in the government	0	2	0	0	0	2	0	0	0	0	0	9	1	0	0
per.politician.headofgovernment: A head of government, a person who rules or commands	per.professionalposition.unspecified: A person in a professional position	0	0	2	0	0	0	2	0	0	0	1	3	4	3	0
per.politician.headofgovernment: A head of government, a person who rules or commands	per.militarypersonnel.unspecified: Someone who serves in the armed forces; a member of a military force	0	1	1	0	0	0	1	1	0	0	0	2	9	8	1
per.politician.headofgovernment: A head of government, a person who rules or commands	per.politician.unspecified: A leader engaged in civil administration	0	2	0	0	0	1	1	0	0	0	2	6	2	0	0
per.professionalposition.unspecified: A person in a professional position	per.unspecified.unspecified: A human person or group of people	0	0	2	0	0	0	2	0	0	0	1	0	5	4	0

Figure 19: Sample Lenient, Strict, and Crowd Semantic Similarity Judgments

Overall, the strict judgment approach tended to result in slightly lower scores than the lenient approach for LDC annotators, as expected. The crowdworker scores were more distributed across the possible scores for each item, but tended to cluster enough to show convergence toward one point on the scale. It was not necessarily the case that the crowdworker “converged” score corresponded to either the lenient or strict LDC annotator score.

When LDC annotators were tasked with tagging entities in a small number of documents using the full DWD rather than the restricted LDC tagset, the results were interesting in several ways. Strict type matching f-scores had an average of 36.95 and a maximum of 44.54 across pairwise comparison of annotators. When similarity rather than strict matching was used, scores went up (average of 64.13, maximum of 66.05). Because there was no closed list of entity types to look for in the data, annotators varied in how many mentions they tagged as entities overall. The differing number of labeled mentions in each version of the document makes mapping annotations for comparison more challenging than with a closed tagset. Not surprisingly, the overall agreement numbers were lower for the open DWD ontology than for a closed tagset given the much larger number of choices available to the annotators. Encouragingly, however, the similarity-based scoring provides a higher level of agreement than the strict match, suggesting that annotators are likely selecting semantically compatible, though not identical, tags much of the time.

5 Conclusions

The LDC data produced for AIDA makes several valuable lasting contributions to the resources available for tackling NLP problems in the spaces of conflicting information, misinformation, complex multi-event hypotheses and claim tracking. In addition to providing data known to contain conflicting accounts of events and annotated for a wide variety of entity, relation, and event types, KB Linking, and the entirely novel Claim Frame task, the data provided are multilingual and multimodal. The annotations cover over 1100 documents and represent over 1 million annotation decisions. This data legacy is summarized in the following table.

Table 30: AIDA Data Legacy

	M9 Pilot	Phase 1	Phase 2	Phase 3
Source documents annotated	423	457	100	203
Mention annotation decisions	44382	247917	81410	187389
Coreference & KB linking decisions	5101	27708	10138	27067
Hypothesis & claim frame decisions	58249	8133	3199	34953
TA1 assessment decisions	12256	16496	n/a	n/a
TA2 assessment decisions	46094	50065	4999	4585
TA3 assessment decisions	n/a	55662	85422	13506
Total decisions	166,505	405,981	185,168	267,500

The data created for AIDA included 15 new corpora, all of which will be published in the LDC catalog in consolidated final versions.

In addition, the work LDC has done for AIDA on similarity-based scoring has already been of use in measuring agreement on annotation in another DARPA program (KAIROS), which also uses the DWD Q nodes as type labels, and the scoring software created for that effort has been distributed to other KAIROS teams involved in the Ontology and Similarity Working Groups who are using it in their efforts to improve similarity scoring.

6 Recommendations

One of the most important concepts throughout AIDA evaluations was “lenient assessment”, which meant that human assessors gave credit to systems for identifying the same informational content, even if the structural representation was somewhat different from the reference annotation. In the final phase of AIDA, with the transition to the use of DWD, considerable program resources on the part of LDC and other AIDA performers were devoted to supporting the ability to automatically score system output based on its semantic similarity to a human reference as well as automatically scoring two different human reference annotations against each other to measure IAA.

Understanding conflicting narratives about the same real-world events – whether during annotation or assessment – requires a holistic understanding of meaning conveyed by the data. Thus far, the work on various algorithms for similarity has focused on scoring pairs of items (concepts) in isolation. While finding a metric that adequately corresponds to human intuitions about the similarity of concepts is an important first step, there are many additional factors to consider in order to compare complex structures such as events, claim frames, or hypotheses.

Multiple human decisions go into annotation of a single event, and each of those decisions affects subsequent decisions. In order to tag a single event, annotators must decide (at least)

1. Is this item taggable? If so, what is its provenance?
2. Is it an event, relation or entity?
3. What is its type or Qnode (possibly multiple - identity, type, base)
4. What are its arguments?
5. Are those arguments taggable and what is the provenance for each?
6. What are the argument roles?
7. What is the type/Qnode for each argument?
8. Are there attributes for any argument?
9. What is the temporal information for the event?
10. Does the event have attributes?
11. How does this event/relation/argument/entity relate to other KEs in the document?

During assessment, human annotators are able to take a holistic view of the “meaning” of the event in order to decide if a system event matches a human reference event. A comparable automatic measure would need to likewise account for the whole event structure as a single unit. Figure 20 illustrates some of the ways in which human assessors account for differences in various aspects of events when making the decision about whether a human and system event pair should be matched.

SYSTEM		REFERENCE
<arg1> caught fire or exploded from <arg2> instrument at <outside the city's courthouse>	Event type mismatch, and Derry isn't mentioned in Sys, but "courthouse" is very salient. ALIGN.	<bombers who may be associated with New IRA & New IRA> detonated or exploded <the bomb used in the attack> explosive device using <arg3> to attack <arg2> target at <Derry courthouse>
<suicide bomber Taimur Abdulwahab al-Abdaly> transported <arg2> in <arg3> from <arg4> place to <there> place	There are two different places Taimour traveled to. Need (unlabeled) doc context to disambiguate. ALIGN to Syria after checking provenance.	<Taimur Abdulwahab al-Abdaly> transported <arg2> in <arg3> from <arg4> place to <Iraq> place
<terrorist> attacked <arg2> using <arg3> at <the Drottningatan area> place	Although target perspective includes attacks in Drottningatan, provenance check reveals that this doc is about a historical attack. DO NOT ALIGN.	<Taimur Abdulwahab al-Abdaly> transported <arg2> in <arg3> from <arg4> place to <Syria> place
		<terrorist> attacked <arg2> using <arg3> at <the Drottningatan area> place

Figure 20: Holistic Event Alignment by Human Assessors

The first event pair in the figure above shows how an assessor can align a fully complete human event with a somewhat impoverished system event, while the second pairing illustrates how assessors sometimes need to make use of document context when systems return ambiguous mentions to fill argument slots. In the third example, events that on the surface look identical are revealed to be entirely different real-world events when document context is examined.

A possible contribution to the resources that would be needed to improve automatic similarity scoring would be annotation of richer event relations. The cross-claim relations in Phase 3 of AIDA are an example of a useful kind of event relation labeling. By identifying claims as supporting or refuting one another, as well as identifying identical and related claims, rich information is provided that could aid in the development of similarity scoring at the claim level, accounting for all aspects of the complex claim object.

Additional event-event relations that could be added to event representations to provide an even richer set of information include explicitly distinguishing strict and fuzzy coreference decisions via the use of full vs partial coreference designations, pro- and anti-causal event relations, enablement relations, and sub-event relations. Even when not explicitly marked, an understanding of such relations may be implicitly used by human assessors when making matching or similarity judgements in a holistic way, and so they might also be useful in developing automatic similarity measures.

7 References

- Dan, V., Paris, B., Donovan, J., Hameleers, M., Roozenbeek, J., van der Linden, S., and von Sikorski, C. (2021). Visual mis-and disinformation, social media, and democracy. *Journalism & Mass Communication Quarterly*, 98(3):641-664.
- DARPA. (2012). Broad Agency Announcement: Deep Exploration and Filtering of Text (DEFT). Defense Advanced Research Projects Agency, DARPA-BAA-12-47.
- DARPA. (2014). Broad Agency Announcement: I2O Low Resource Languages for Emergent Incidents (LORELEI). Defense Advanced Research Projects Agency, DARPA-BAA-15-04.
- DARPA. (2017). Active Interpretation of Disparate Alternatives (AIDA). Defense Advanced Research Projects Agency, DARPA BAA HR001117S0026.
- DARPA (2019). Broad Agency Announcement : Knowledge-directed Artificial Intelligence Reasoning Over Schemas (*KAIROS*). Defense Advanced Research Projects Agency BAA HR001119S0014.
- Grabar, N. and Hamon, T. (2019). WikiWars-UA: Ukrainian corpus annotated with temporal expressions. In *Proceedings of Computational Linguistics and Intelligent Systems (COLINS 2019)*, Kharkiv, Ukraine, April
- Grabar, N., Kanishcheva, O., and Hamon, T. (2018). Multilingual aligned corpus with Ukrainian as the target language. In *Proceedings of SlaviCorp 2018*, Prague, Czech Republic, September
- Gottlieb, M., and Dyer, S. (2020). Information and Disinformation: Social Media in the COVID-19 Crisis. *Academic Emergency Medicine*, 27(7):640-641.
- Han, Q., John, M., Kurzhals, K., Messner, J., and Ertl, T. (2018). Visual interactive labeling of large multimedia news corpora. In *Proceedings of Leipzig Symposium on Visualization in Applications (LEVIA'18)*, Leipzig, Germany, October.
- Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick C. L., and Dollár, P. (2015). Microsoft COCO: Common objects in context. <https://arxiv.org/abs/1405.0312v3>, accessed 1/13/2022.
- NIST. (2019). Streaming Multimedia Knowledge Base Population (SM-KBP) 2019. Text Analysis Conference. <https://tac.nist.gov/2019/SM-KBP/index.html>
- Rossetto, L., Giangreco, I. and Schuldt, H. "Cineast: A Multi-feature Sketch-Based Video Retrieval Engine," *2014 IEEE International Symposium on Multimedia*, Taichung, Taiwan, 2014, pp. 18-23, doi: 10.1109/ISM.2014.38.
- SciPy. (2016). https://docs.scipy.org/doc/scipy-0.18.1/reference/generated/scipy.optimize.linear_sum_assignment.html. Accessed 2/14/23.
- Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., and Liu, H. (2020). Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6), e1385.

- Shvedova, M., (2020). The general regionally annotated corpus of Ukrainian (GRAC, uacorporus.org): Architecture and functionality. In *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020)*. Lviv, Ukraine, April, pages 489-506.
- Sitchinava, D. (2012). Parallel corpora within the Russian National Corpus. *Philological Studies*, 63: 271-278.
- Szostek, J. (2018). Nothing is true? The credibility of news and conflicting narratives during “information war” in Ukraine. *The International Journal of Press/Politics*, 23(1):116–135.
- Tracey, J. and Strassel, S. (2020). Basic language resources for 31 languages (plus English): The LORELEI representative and incident language packs. In *Proceedings of the 12th Edition of the Language Resources and Evaluation Conference (LREC 2020)*, pages 277–284, May. European Language Resource Association (ELRA).
- Vrandečić, D. and Krötzsch, M. (2014). "Wikidata: a free collaborative knowledgebase." *Communications of the ACM* 57.10: 78-85

8 Appendix A: Corpora Produced

Recipients	Catalog ID	Title
AIDA	LDC2018E01	AIDA Scenario 1 - Seedling Corpus V2.0
AIDA	LDC2018E45	AIDA Scenario 1 - Seedling Annotation V6.0
AIDA	LDC2018E52	AIDA Scenario 1 - Seedling Corpus Part 2 V2
AIDA	LDC2018E53	AIDA Scenario 1 - Seedling Background Corpus Non-Eval
AIDA	LDC2018E62	AIDA Month 9 Pilot Evaluation Corpus
AIDA	LDC2018E63	AIDA Scenario 1 - Seedling Corpus Training Data Video Segmentation V2.0
AIDA	LDC2018E64	AIDA Scenario 1 - Seedling Background Corpus Non-Eval Video Segmentation V1.0
AIDA	LDC2018E73	AIDA Ukrainian Broadcast and Telephone Speech Transcripts V2.0
AIDA	LDC2018E74	AIDA Ukrainian Broadcast and Telephone Speech Audio V1.0
AIDA	LDC2018E76	AIDA Month 9 Pilot Eval Annotation - Unsequestered V1.0
Limited	LDC2018R26	AIDA Month 9 Pilot Eval Annotation
AIDA	LDC2019E04	AIDA Phase 1 Evaluation Practice Topic Source Data V2.0
AIDA	LDC2019E05	AIDA Scenario 1 Re-Annotation Format Sample
AIDA	LDC2019E07	AIDA Phase 1 Evaluation Practice Topic Annotations V8.0
AIDA	LDC2019E42	AIDA Phase 1 Evaluation Source Data
AIDA	LDC2019E43	AIDA Phase 1 Evaluation Reference Knowledge Base
AIDA	LDC2019E44	AIDA Phase 1 Practice Topics Reference Knowledge Base
AIDA	LDC2019E77	AIDA Phase 1 Evaluation Topic Annotations - Unsequestered V2.0
AIDA	LDC2019E80	AIDA Phase 1 Assessment Results - Unsequestered
Limited	LDC2019R05	AIDA Month 9 Pilot Eval Assessment Results V2.0
Limited	LDC2019R07	AIDA Phase 1 Annotation for PAR V4.0
Limited	LDC2019R14	AIDA Phase 1 Evaluation Topic Annotations V1.2
Limited	LDC2019R20	AIDA Phase 1 Evaluation Source Data for PAR
Limited	LDC2019R21	AIDA Phase 1 Eval Annotation for PAR V1.0

Limited	LDC2019R30	AIDA Phase 1 Assessment Results V6.0
AIDA	LDC2020E11	AIDA Phase 2 Practice Topic Source Data
AIDA	LDC2020E22	AIDA Reannotated M18 Exhaustive Text Sample Annotation V1.1
AIDA	LDC2020E27	AIDA Phase 2 Practice Topics Reference Knowledge Base V1.1
AIDA	LDC2020E29	AIDA Phase 2 Practice Topic Annotation V3
AIDA	LDC2020E32	AIDA TAC SM-KBP 2020 Evaluation Source Data
AIDA	LDC2020E37	AIDA Phase 2 Evaluation Topics Reference Knowledge Base - Unsequestered
AIDA	LDC2020E38	AIDA Phase 2 Evaluation SINS and Perspectives - Unsequestered
Limited	LDC2020R08	AIDA Phase 2 Eval Topic Source Data Superset for PAR
Limited	LDC2020R17	AIDA Phase 2 Evaluation Source Data
Limited	LDC2020R18	AIDA Phase 2 Evaluation Preliminary Evaluation Text Source for NIST
Limited	LDC2020R19	AIDA Phase 2 Evaluation Annotation V2.0
Limited	LDC2020R20	AIDA Phase 2 Evaluation Topics Reference Knowledge Base
Limited	LDC2020R28	AIDA Phase 2 Assessment Results V3.1
AIDA	LDC2021E04	AIDA Phase 3 Hackathon V1.0
AIDA	LDC2021E11	AIDA Phase 3 Practice Topic Source Data V2.0
AIDA	LDC2021E15	AIDA Phase 3 TA1 Practice Topic Annotation V6
AIDA	LDC2021E16	AIDA Phase 3 TA3 Practice Topic Annotation
AIDA	LDC2021E17	AIDA Phase 3 TA3 Practice Topic Dry Run Query Claim Frames
Limited	LDC2021R02	AIDA Phase 2 TA3 Rerun Assessment Results V4
Limited	LDC2021R06	AIDA Phase 3 Preview Data for MITRE
Limited	LDC2021R08	AIDA TA1 Practice Topic Annotation Source Data for PAR
Limited	LDC2021R13	AIDA Phase 3 TA1 Eval Topic Annotation Source Data for PAR
AIDA	LDC2022E06	AIDA Ukrainian Broadcast and Telephone Speech Audio and Transcripts
AIDA	LDC2022E10	AIDA Phase 3 Assessment Results - Unsequestered
Limited	LDC2022R01	AIDA Phase 3 TA3 Evaluation Topic Query Claim Frames V3
Limited	LDC2022R02	AIDA Phase 3 Evaluation Source Data

Limited	LDC2022R09	AIDA Phase 3 TA1 Practice Topic Annotation for NIST
Limited	LDC2022R11	AIDA Phase 3 TA3 Evaluation Topic Annotation
Limited	LDC2022R12	AIDA Phase 3 Assessment Results
Limited	LDC2022R13	AIDA Phase 3 TA1 Evaluation Topic Annotation

9 List of Acronyms and Abbreviations

Acronym	Definition
AIDA	Active Interpretation of Disparate Alternatives
ARG	Argument; a variable within a claim
BAA	Broad Agency Announcement; proposals solicitation
CIA	Central Intelligence Agency; foreign intelligence service of the U.S.
COVID	COrona VIRus Disease
COVID-19	Disease caused by SARS CoV-2 virus
DARPA	Defense Advanced Research Projects Agency
DEFT	Deep Exploration and Filtering of Text
DWD	DARPA Wikidata; a snapshot of Wikipedia customized for use in the AIDA program
ENG	English
EVAL	Evaluation
FSB	Federal Security Service; Russian intelligence service
GPE	Geo-Political Entity type
GNB	Bolivian National Guard
GRAC	General Regionally Annotated Corpus
H#	Hypothesis
HTML	HyperText Markup Language. Standard markup language for web browser display
IAA	Inter-annotator agreement
ID	IDentifier
K	Thousand; e.g. 19K = 19,000
KAIROS	Knowledge-directed Artificial Intelligence Reasoning Over Schemas
KB	Knowledge base
KE	Knowledge element; events and relations, as well as their arguments

LDC	Linguistic Data Consortium
LORELEI	Low Resource Languages for Emergent Incidents
LTF	Logical Text Format
MD5SUM	A computer program that calculates and verifies 128-bit MD5 hashes. It functions as a compact digital fingerprint of a file, and is used to verify file integrity at either end of a file transfer or some other process that might compromise a file
MH17	Malaysian Airlines Flight 17
MITRE	American NPO managing federally funded research and development centers (FFRDCs)
MP#	File extension - MPEG (Moving Picture Experts Group; video or video and audio file); mp4 is current standard
NIL	Not meaningful or salient for given purposes
NLP	Natural Language Processing
NPI	Non-Pharmaceutical Interventions
ORG	Organization entity type
PAR	PAR Government Systems
PER	Person entity type
PPE	Personal Protective Equipment (e.g., face mask)
PRC	People's Republic of China
PSM	Primary Structural Markup
QC	Quality Control
Q NODE	Query node; DWD element identifier, e.g., China (Q148)
RSD	Raw Source Data
RUS	Russian
SARS-CoV-2	Virus that causes COVID-19 disease
SEG	Segment
SIN	Statement of Information Need
SPA	Spanish

SSU	Security Service of Ukraine; Ukrainian intelligence service
TA#	AIDA Program Technical Area (e.g. TA1, TA2, TA3 ...)
TAC KBP	Text Analysis Conference Knowledge Base Population
UID	Universal IDentifier
UKR	Ukraine/Ukrainian
URL	Universal Resource Locator; internet address
X VARIABLE	Field in a claim template to be filled with the query's search object
XML	Extensible Markup Language; defines a set of rules for encoding documents in a format that is both human-readable and machine-readable