# Strategic Plan, Roadmap and Requirements Overview for Structured Representation of Digital Investigation Information

## Strategic Plan

**Goal:**

Develop, define and enable a standardized information representation for the capture, analysis and exchange needs of the cyber-centric digital investigation domain.

**Targeted use case strategic mission scope:**

Digital forensics, cyber investigation, incident response, and counter terrorism.

**Approach:**

Collaborate among a diverse and open community of stakeholders to formally define a practical information representation for the digital investigation domain while adhering to a core set of principles:

1. Pursue, value and leverage the experience-based input of domain experts.
2. Pursue, value and leverage the experience-based input of non-domain experts in information representation
3. Pursue a formal model-based specification approach (conceptual model -> OWL ontology + textual specification document -> serialization binding specifications -> reference serialization implementations (e.g. JSON-LD)) to minimize ambiguity and maximize potential for automation and consistency.
4. Strive for a representation that is expressive, flexible, consistent and as simple as possible
5. Enable interoperability between tools, between investigations, between cyber defense sub-domains (digital investigations, malware analysis, attribution, threat detection, etc.), and between organizations
6. Recognize that Digital Investigation is only one facet of a broader cyber defense domain and that DI information is most effective when placed within the broader context.

In recognition of principle #6 and supporting #1-#5, this effort will pursue an information representation for digital investigations that inherently recognizes and is designed for its place within the broader domain of cyber defense. Taking into account principle #3, this effort will focus on defining and specifying a digital investigation ontology as one particular component of a broader Uniform Cyber Ontology (UCO). The purpose of the UCO is to provide a unifying framework to align and integrate disparate forms of cyber information from varying focused sub-domains (e.g. digital investigations) and to provide foundational concepts and constructs for representing these disparate forms of cyber information in a consistent, automatable and interoperable fashion. Each sub-domain is best defined, evolved and managed by the community of experts focused on that area. While this effort is primarily focused on the sub-domain of digital investigation (uco-di) it is foreseen that initial definitions will also likely be necessary for several other related and relevant sub-domains include foundational concepts and constructs (uco-core), cyber observables (uco-observable) given recent troubling developments in the OASIS CTI TC plans for CybOX 3.0 and onward, actors (uco-actor), victims (uco-victim), actions (uco-action) and potentially others. This effort will seek out experts in these related domains to contribute to the validation and evolution of the initial ontology specifications produced.

**Overall deliverable objectives:**

* Strategic Plan document
  + Approach
  + Use Cases
  + Requirements
  + Roadmap
* Community communications plan including collaboration mechanisms and platforms (email list, github project, etc)
* UML conceptual model of digital investigation information representation needs (digital investigation-centric as well as any related required domains)
* Formal OWL ontology derived from and aligned to the UML conceptual model
* Text-based data model specification (with UML and OWL adornment)
* JSON-LD serialization binding specification
* Serialization-specific framing definitions for JSON and XML
* Basic API base (Python?, JAVA?, Go?, other?) for input and output of serialized content in support of tool implementers
* Develop mappings and bridging ontologies for relevant related representations (e.g. DOMEX)
* Implementation in plaso
* Implementation in 2-3 other tools (e.g. Sleuthkit)
* Appropriate documentation to be determined

## Roadmap

**Immediate-term task goals:**

* Establish a roadmap and initial, draft, high-level informal ontology (based on DFAX but with several modifications) that all participants can agree on
* Establish effort nomenclature, branding and communication plan
* Identify and recruit other key stakeholders into the effort community

**Near-term task goals:**

* Develop initial draft UML conceptual model derived from consensus initial, draft, high-level informal ontology
* Test OWL and JSON-LD generation from UML conceptual model
* Develop an initial proof-of-concept serialization in JSON. This must be tool agnostic, but implemented in plaso as a demonstration, and must also be aligned with the existing NFI JSON.
* Develop initial proof-of-concept API base library.
* Continue ongoing mapping efforts to relevant related representations
* Complete initial draft implementation in plaso

**Mid-term task goals:**

* Iterate and refine UML conceptual model based on continuing community input/feedback and experimentation
  + Focus on validating and fleshing out uco-di as well as extending uco-observable item/facet coverage
* Develop and refine formal OWL ontology
* Develop initial draft of text-based data model specification
* Develop initial JSON-LD serialization implementation and initial draft of JSON-LD serialization binding specification
  + Develop initial framing rules for JSON serialization
* Complete initial draft implementations in 2-3 other tools?
* Develop strategic plan for alignment and integration with broader threat-centric cyber defense domain and ecosystem

**Long-term task goals:**

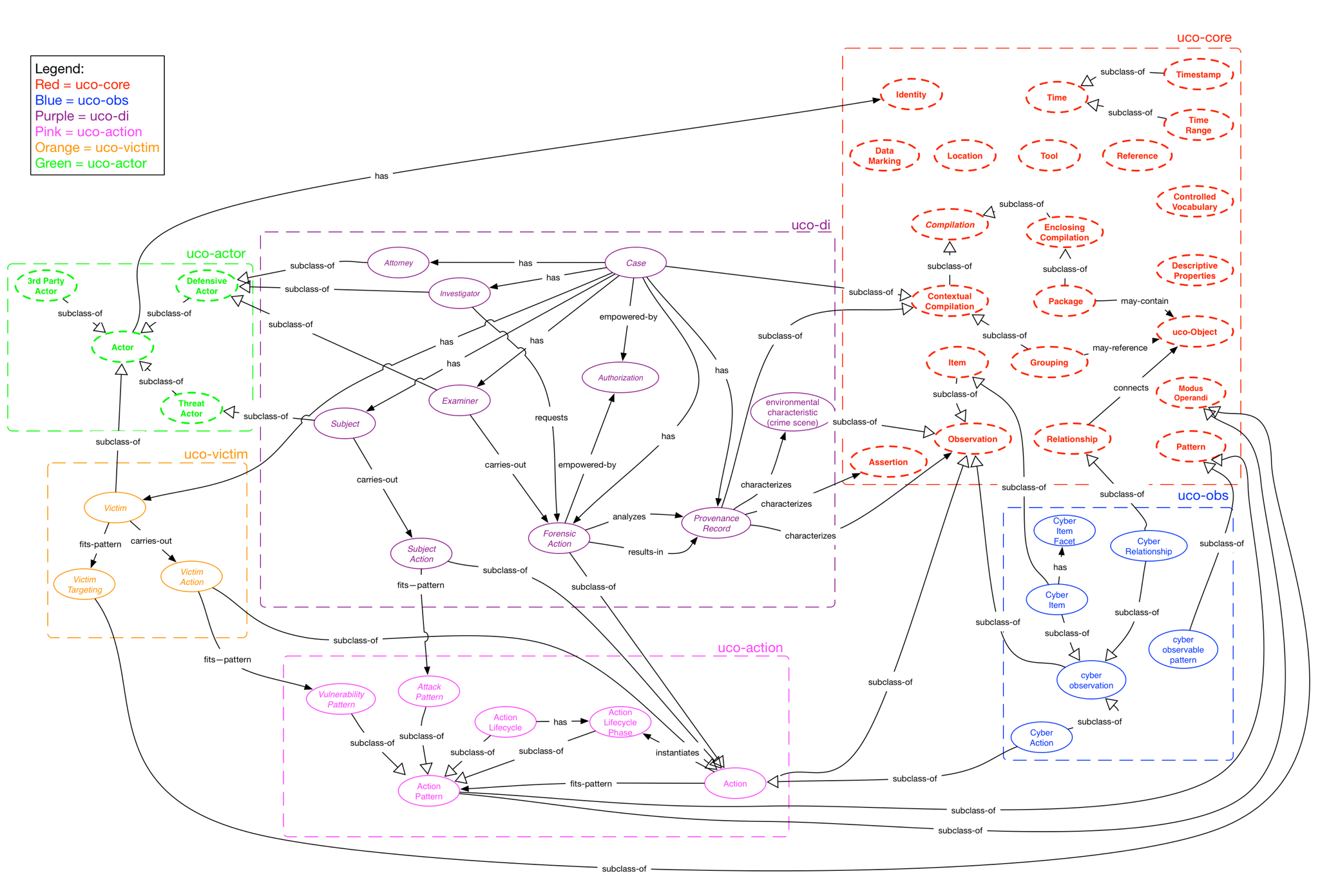
* Iterate and refine UML conceptual model, OWL ontology, text-based data model specification, JSON-LD serialization binding specification and related implementations, and documentation based on continuing community input/feedback and experimentation
* Iterate, refine and improve API base libraries based on community needs
* Iterate and refine tool implementations based on community needs and refinements in underlying data model
* Refine and execute against strategic plan for alignment and integration with broader threat-centric cyber defense domain and ecosystem

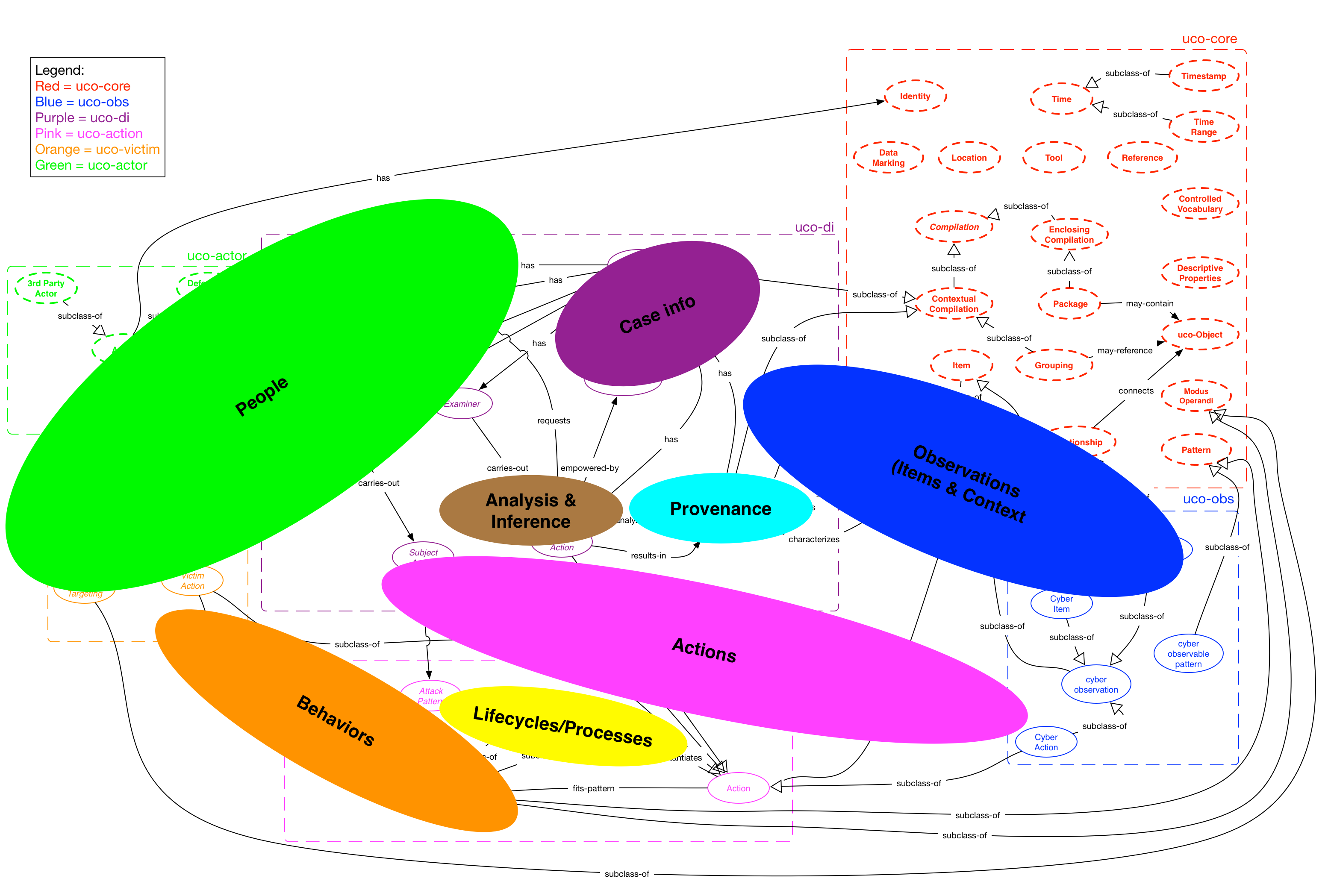
## Vocabulary:

* Uniform Cyber Ontology (uco): *<definition>*
* Uniform Cyber Ontology – Core (uco-core): *<definition>*
  + UCO-Object: Base object defining the minimal core set of properties to act as a consistent, unifying and interoperable foundation for all explicit and interrelateable content objects within the Uniform Cyber Ontology (uco).
  + Descriptive Properties: *<definition>*
  + Relationship: An association or link between two uco objects.
  + Controlled Vocabulary: *<definition>*
  + Reference: *<definition>*
  + Data Marking: *<definition>*
  + Tool: Characteristics of a tool used in a cyber context potentially including its usage environment and configuration characteristics.
  + Location: Characterization of a physical geolocation.
  + Timestamp: *<definition>*
  + Time Range: *<definition>*
  + Identity: Characterization of the identifying properties of an individual or organization.
  + Observation: *<definition>*
  + Item: Physical or digital objects or entities, such as a mobile device, a file extracted from a device, an email address extracted from a file.
  + Assertion: Statement of opinion based on inference, reasoning or analytic interpretation of factual observations and other relevant context.
  + Pattern: *<definition>*
  + Package: *<definition>*
  + Grouping: An assortment of items that are logically grouped together. For instance, a MESSAGES are logically grouped into a THREAD.
* Uniform Cyber Ontology – Observable (uco-obs): *<definition>*
  + Cyber Observation: *<definition>*
  + Cyber Item: *<definition>*
  + Cyber Item Facet: Contextually related characteristics of an item. For example, a file can have multiple facets (file, file system, picture, JPG, thumbnail) each with its own set of properties.
  + Cyber Action: *<definition>*
  + Cyber Relationship: An association or link between two uco-obs objects.
* Uniform Cyber Ontology – Action (uco-action): *<definition>*
  + Action: Some observed activity caused by a person or system, which may be associated with item(s). For instance, execution of malware can result in various files, Registry entries, logs, etc. Actions can be an adjunction (object added to the crime scene, matter transferred, malware stored on a computer), a suppression (object stolen, file deleted), a transformation (misplaced/displaced object in a crime scene, file moved), a change of state (modified access rights of a file).
  + Action Pattern: *<definition>*
  + Action Lifecycle: *<definition>*
  + Action Lifecycle Phase: *<definition>*
  + Attack Pattern: *<definition>*
  + Vulnerability Pattern: *<definition>*
* Uniform Cyber Ontology – Digital Investigation (uco-di):
  + Case: *<definition>*
  + Forensic Action: *<definition>*
  + Provenance: Where an item originated or was found, as well as how an item was handled after it was found.
  + Provenance Record: *<definition>*
  + Environmental Characteristic: *<definition>*
  + Authorization: *<definition>*
  + Attorney: *<definition>*
  + Investigator: *<definition>*
  + Examiner: *<definition>*
* Uniform Cyber Ontology – Victim (uco-victim):
  + Victim: The individual or organization that is/was the target of malicious action.
  + Victim Action: An action taken by a victim.
  + Victim Targeting: Pattern of identifying characteristics describing what sort of victims are/were targeted by a particular malicious actor or as part of a particular malicious activity.

## Types of information to represent for the digital investigation domain:

* Case & Evidential Items
  + Case details
  + Observations (Items, Actions, and Relationships)
  + Tool details
  + Output / results
* Provenance\_Records & Forensic\_Actions
  + Where an evidential item originated or was found, as well as how an item was handled after it was found (who did what to evidence where, when, and how?)
* Human Actors (Subject, Victim, Examiner, Investigator, Attorney, etc.)
* Forensic Processes (Lifecycle)
* Offender & Victim Processes (Lifecycle)
* Offender & Victim Behaviors (Action Patterns)
* Inferences/Assertions (human reasoning or automated analytics)

Draft Ontology Overview



## Digital Investigation (uco-di) Representation Requirements High-level Overview

**Requirement 1**: Represent multi-faceted extracted items and inter-relationships

Proposal:

1. Define base FILE item, and then define facet extensions for various file types (picture, audio, video, database, MS office, PDF, LNK, Jumplist).
2. Define base MESSAGE item, and then define facet extensions for various message types (sms, email, chat).
3. Define relationships between item, e.g., SQLite “mail.db” FILE contains a “Yahoo! Mail” MESSAGE and a blob FILE attached to the MESSAGE.
4. Represent other facets such as language using facets.

**Requirement 2**: Represent tools and system environment

Proposal:

1. Represent a specific version/configuration of tool or system as a distinct entity, with its own unique identifier.
2. Capture provenance, i.e., each instance of a tool being run and the specific system used to perform the operation.

**Requirement 3**: Have logical Groupings/Assemblages of objects including extracted items

Proposal:

1. Define logical groups, including:
   * The Actor grouping/assemblage logically groups information about a person, including name, aliases, email addresses, age, etc.
   * The Thread grouping/assemblage logically groups message that are part of the same conversation
   * The Narrative grouping/assemblage logically groups information that represent investigation results or outcomes, including a timeline or link diagram.

**Requirement 4**: Represent Actions associated with extracted items, including time and location dimensions.

Proposal:

1. Define Action to apply to any item.

**Requirement 5**: Maintain provenance details for original, extracted, and processed items

Proposal:

1. Define Provenance\_Record, including who did what, where, when, how, and sometimes why.

**Requirement 6**: Represent Forensic\_Actions, both automated and human

Proposal:

1. Define Forensic\_Action, with input either from original evidential item or Provenance\_Record. Output must have an associated Provenance\_Record.
2. Distinguish between Forensic\_Actions that are purely fact-based observations and Forensic Actions that use human reasoning (abductive, deductive, inductive) whether human or automated to reach analytic assertions. For the output of human reasoning, assign a confidence such as a likelihood ratio or a scale (low, medium, high confidence).

**Requirement 7**: Represent Actions of offender(s), victim(s), and system

Proposal:

1. Define Actions associated with item, such as Offender sent a “Yahoo! Mail” MESSAGE to Victim.

**Requirement 8**: Represent behaviors of offender(s), victim(s), and system

Proposal:

1. Define Action Patterns to capture behaviors.

**Requirement 9**: Represent forensic process lifecycle

Proposal:

1. Define Action\_Lifecycle that can be used to represent forensic process lifecycle (e.g., evidence processing, evidence analysis).

**Requirement 10**: Represent offender process lifecycle

Proposal:

1. Define Action\_Lifecycle that can be used to represent offender process lifecycle (e.g., intruder breaking into a network, sexual predator grooming victim.

**Requirement 11**: Represent victim process lifecycle

Proposal:

1. Define Action\_Lifecycle that can be used to represent victim lifecycle.

**Requirement 12:** Represent what language(s) are within digital files/messages

**Requirement 13:** Represent data markings (classification, handling, etc) on content

## Open Questions to resolve:

* Identity and Role
  + Required: Ability to characterize identity
  + Required: Ability to associate an identity with a particular role within a particular context (Fred is a victim in one context and a subject in another and an examiner in another)
  + Proposed approach: Leverage role-based objects (derived from uco-role) to represent roles played and associate those role objects with the relevant context where the role was played, and then define relationships between these role objects and identity objects for the party (individual or organization) who played that role
  + Are there role specific properties for any of the di roles (attorney, investigator, examiner, subject)?
  + Are there role specific actions and relationships?
  + What structure to use for identity characterization?
  + Proposed approach: Start with CIQ, convert to JSON, identify opportunities to simplify without losing necessary expressivity/flexibility, utilize a facet approach for properties
* Action
  + How best to abstract action such that it can support human actions, process actions, cyber actions, etc.?
  + Requirement: Ability to express both abstract actions and detailed actions
  + Requirement: Ability to express relationships between actions
    - inter-action temporal relationships, ordinality, etc.
  + Requirement: Ability to express relationships between actions and items
    - Initiating, affecting, utilizing, resulting
    - What properties are necessary?
  + Requirement: Ability to express composite actions made up of other actions
  + Requirement: Ability to express action patterns
    - What properties are necessary?
  + Requirement: Ability to express action lifecycles
    - What properties are necessary?
* Structuring of items and facets
  + Requirement: Ability to express details of cyber observable items
  + Requirement: Flexibility in characterizing details of different facets of cyber observable items
  + How best to handle facets (single level (one general item type that can contain any number of facets) vs multi-level (multiple item sub-types with sets of facets appropriate to them)?
    - Proposed approach: Single level (one general item type that can contain any number of facets)
      1. Provides better simplicity, flexibility and extensibility
  + How best to address item/facet decompositions (facets within facets) with relationships?
* Non-cyber observations
  + Requirement: Ability to express details of non-cyber observations (environmental details of crime scenes)
  + What level of generic Observation and Item constructs are needed in uco-core?
  + What level of detail is necessary within the uco-di to specifically express details of investigation-centric non-cyber observations?
  + Proposed approach: Leverage general item/facet approach with facets defined for non-cyber characteristics
* Data markings
  + Requirement: Ability to express data markings (sharing, handling, acting, etc.) on uco data
  + Are object level markings needed?
  + Are field level markings needed?
  + Is there a requirement for default marking models?
    - What would they be?
* Tool
  + Requirement: Ability to express details of tools
    - Requirement: Ability to express details of forensic analysis tools including their nature, configuration and environmental characteristics of their use
    - Requirement: Ability to express details of general tools including their nature, configuration and environmental characteristics of their use
    - Requirement: Ability to express details of malicious tools including their nature, configuration and environmental characteristics of their use
  + What level of detail needed?
  + Need to distinguish between beneficial, benign and malicious tools?
  + Proposed approach: Leverage role-based approach similar to that proposed for identity/roles where:
    - A general Tool construct in would implicitly be considered neutral
    - Malicious tools would use a Malicious Tool construct that inherits from Tool and from Malicious Role
    - Forensic Analysis tools would use a Benevolent Tool construct that inherits from Tool and Benevolent Role
* Location
  + Requirement: Ability to express explicit location details
  + What level of detail needed?
  + Any particular model necessary to support?
  + Proposed approach: Leverage refactoring of CIQ from Identity, call it location not address (as it is done in CIQ), utilize a facet approach (allows flexible specification and use of facets for whatever coordinate system (lat-long, GPS, KML, etc) is desired)
* Patterns
  + How best to structure the abstract pattern base?
  + How best to support the inherent differential between cyber observable patterns and human action patterns?
  + Distinguishing between and supporting patterns for conveyance of abstraction vs conveyance of filter/query?
  + Structure vs grammar?
* How best to represent analytic (inference, reasoning, heuristics, etc) assertions?
  + What level of structure needed for expressing evidentiary support for assertion?
* How to handle versioning of content?