(Note: this is an abstract for a presentation at a conference. There is no associated paper.)

Discovery of digital forensic dataset characteristics with CASE-Corpora

The digital forensics community has generated training and reference data over the course of decades. However, significant challenges persist today in the usage pipeline for that data, from research problem formulation, through discovery of applicable shared data, through local processing and analysis. The problems include classic afflictions of Internet resources such as link rot and maintainer departure. Greater challenges remain in covering the gap between research questions in natural language, and the structured metadata in available dataset annotations. A dataset that describes a picture sent between two accounts in a chat app entails many levels of pattern abstraction, and should be discoverable whether a user is searching datasets for, e.g., that app, or for any instance of picture transmission. Yet, disparate datasets today lack a unifying language which can be aggregated and queried in one location.

We present CASE-Corpora, a community index of available forensic reference and training datasets. CASE-Corpora aggregates dataset descriptions into a single ontological knowledge graph. Several ontologies are exercised to enable representation of a dataset, its downloadable resources, and how those resources may have migrated between hosts over time. These are represented in the Data Catalog (DCAT) and Provenance (PROV) Ontologies. However, these ontologies intentionally abstain from rich representation in focused domains such as digital forensics. They can express metadata of where to find data, but not why a forensic analyst would want to find it.

CASE-Corpora describes forensically-relevant qualities of datasets by way of the Cyber-investigation Analysis and Standard Expression (CASE) Ontology and the Unified Cyber Ontology (UCO). CASE-Corpora uses all of the above ontologies to describe not only what one would download, but what devices, actions, and environmental captures went into the download; the hashes of both downloadable resources and what would be extracted for analysis, verifying chain of custody; and, where available, ground truth for analytic results' cross-verification. CASE and UCO exercise ontology-level interoperability to expertise and practice with domain-agnostic dataset discovery and provenance review to apply to finely-specified forensic detail.

This presentation introduces CASE-Corpora and its support for curation and growth by the community. Data review mechanisms assist with ensuring conformant usage of the employed ontologies. Queries written for the corpora show the richness of what the community has already developed for research, such as what devices have been involved in any dataset. Our experience with developing CASE has repeatedly and consistently shown immense value in defining questions and encoding them as queries. For CASE, this has grown the ontology. For the community, CASE-Corpora can grow our collective knowledge on discoverable and testable patterns in data that many in the community have put effort into making available. This presentation will improve ontological competency, analysis interoperability, and data discovery for the community.

This presentation will cover experiences growing CASE-Corpora since an earlier edition presented at the Digital Forensics Research Workshop in July of 2022. Experiences will have been shaped by encountering potential errors in dataset chain of custody, as well as the CASE version 1.0.0 release that initialized backwards compatibility guarantees until the version 2.0.0 release.