Giovanni Cozzolino

Tutor: Antonino Mazzeo

co-Tutor: Flora Amato

XXXI Cycle - III year presentation

A semantic methodology for (un)structured digital evidences analysis

Università degli Studi di Napoli
Federico II

PhD Candidate

Giovanni Cozzolino

- Graduation: MSC in Computer Engineering
- Fellowship: DIETI Grant
- Research Field: Digital Forensics, Knowledge management, Data Integration
- Projects:
  - CREA - CONFLICT RESOLUTION WITH EQUITATIVE ALGORITHMS
  - CRUI – Conferenza dei Rettori delle Università Italiane, "Strumenti e funzionalità del Processo Civile e Penale telematico, profili di sicurezza dei sistemi informativi in uso presso il Ministero della Giustizia e gli uffici Giudiziari"

PhD Candidate

Giovanni Cozzolino

Student: **Giovanni Cozzolino**
giovanni.cozzolino@unina.it

Cycle: **XXXI**

Tutor: **Antonino Mazzeo** antonino.mazzeo@unina.it

Co-Tutor: **Flora Amato** flora.amato@unina.it

	Credits year 1								Credits year 2								Credits year 3									
		1	2	3	4	5	6			1	2	3	4	5	6			1	2	3	4	5	6			
	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Total	Check
Modules	20		6		6		6	18	12	0	3	3	3	0	0	9	6	0	6	0	0	0	0	6	33	30-70
Seminars	5							0	10	0,8	0,2	1,8	3,5	0	1,1	7,4	5	0,4	0,6	0	1,2	0,3	0,4	2,9	10,3	10-30
Research	35	6	5	8	6	4	6	35	45	8	7	8	6	8	8	45	60	10	10	10	10	10	10	60	140	80-140
	60	6	11	8	12	4	12	53	67	8,8	10	13	13	8	9,1	61	71	10	17	10	11	10	10	69	183	180

Context overview

Information technologies are pervasive

Information Technology (IT) systems are involved in almost all daily activities

- Computers
- Mobile devices
- Embedded systems
- Networks
- Databases

Each activity performed on a digital device routinely leaves some kind of digital fingerprint.

Context overview

Digital traces are valuable

Nowadays, the world's most valuable resource are digital data.

Valuable information

Business, Health, Communication and Social, Finance, Mobile, Security, etc..

Business purposes

Patents and trademarks, Marketing,

Criminal behaviors

Spam, malware, frauds, stalking, terrorism, etc..

The examination, interpretation and reconstruction of digital fingerprint in a digital environment falls within the realm of Digital Forensics.

Digital Forensic

A support for criminal or intrusion investigations

Digital forensic investigators retrieve and process digital evidences, found on computers and networks, ensuring their admissibility in a legal context.

Main phases:

- Identification

It consists in the choice of the devices or the systems to be examined.

Digital Forensic

A support for criminal or intrusion investigations

Digital forensic investigators retrieve and process digital evidences, found on computers and networks, ensuring their admissibility in a legal context.

Main phases:

- Identification
- Preservation

It must guarantee the isolation of the examined evidence from external agents and, in general, the outside world.

Digital Forensic

A support for criminal or intrusion investigations

Digital forensic investigators retrieve and process digital evidences, found on computers and networks, ensuring their admissibility in a legal context.

Main phases:

- Identification
- Preservation
- Acquisition

It involves the use of specific tool in order to acquire and preserve the information

Digital Forensic

A support for criminal or intrusion investigations

Digital forensic investigators retrieve and process digital evidences, found on computers and networks, ensuring their admissibility in a legal context.

Main phases:

- Identification
- Preservation
- Acquisition
- Analysis and Correlation

It strictly depends on the scenario and on requirements for searches.

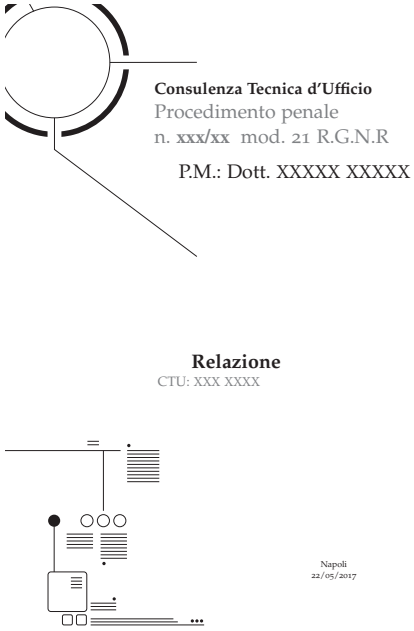
Digital Forensic

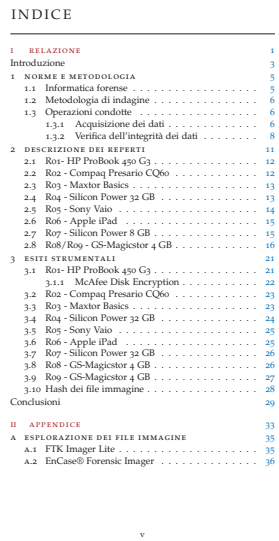
A support for criminal or intrusion investigations

Digital forensic investigators retrieve and process digital evidences, found on computers and networks, ensuring their admissibility in a legal context.

Main phases:

- Identification
- Preservation
- Acquisition
- Analysis and Correlation
- Documentation





It formalizes the examiners work, describing all the operation conducted.

Motivation

Digital Forensics investigations are becoming challenging

Large amount of data

- The size of memory devices increases constantly.
  - Hard disk, SAN, smartphone…
- Live acquisition of streams.
  - SN analysis, real-time or surveillance systems…

Heterogeneity of data

- Many tools for the acquisition process.
  - produce different representation format.
- Relevant information can be structured or not
  - records, log files, XML…
  - text files, PDF, images…

The interpretation of such information allow for the contextualization of digital evidences, promoting them as clues.

XML-based approach (1)

Digital Evidence Exchange (DEX)

A standard and open format for digital evidence provenance.
A DEX file is an XML-based description of objects of interest - volumes, files, JPEGs, etc. - found in a raw image file.

It is designed both for:

- description and comparison of particular pieces of evidence
- tool interoperability and validation

Investigators can use DEX through:

- A public an open-source library that supports creation and comparison of DEX files
- A set of wrappers for several existing forensic tools

XML-based approach (2)

Digital forensics XML (DFXML)

- DFXML supports the representation, by a family of XML elements, of the following kinds of forensic data:
  - Metadata describing the source disk image, file, or other input.
  - Information about the forensic tool used for processing data.
  - The state of the computer on which the processing was performed (e.g., the name of the computer) and Operating-system-specific information useful for forensic analysis.
  - The evidence that was extracted, how it was extracted, and where it was physically located.
- There are toolkits for reading and generating DFXML.
- A lot of forensic tools or distributions support it.

RDF-based approach

Advanced Forensic Format (AFF)

The AFF is a format used to represent forensic images. It offers compression and metadata storing within the image.
Latest version of this format is AFF4. It was designed following some basic assumptions:

- AFF4 Objects are simply entities about which statements are made. AFF4 objects have a globally unique name (called a URN).
- AFF4 uses RDF to model the statement about the object in the form of (subject, predicate, value).
- AFF4 statements are stored in a Resolver which manages the information model.

Ontology-based approach

Forensics for Rich Events (FORE)

FORE (Forensics for Rich Events) architecture is composed of:

- a forensic ontology, based on two main concepts that represent tangible objects and their state changes over time
- an event log parser
- a custom defined rule language, FR3
  - Rules expressed in FR3 language are evaluated against the knowledge base in order to add new causality links between instances

Idea

Use Semantic-Web technologies to represent domain concepts and data through a reliable and standardized format.

Main potential advantages of such an approach regard:

- Extensibility: compatibility with data model of forensic tools input and output;

Idea

Use Semantic-Web technologies to represent domain concepts and data through a reliable and standardized format.

Main potential advantages of such an approach regard:

- Extensibility: compatibility with data model of forensic tools input and output;
- Flexibility: through custom ontologies definition;

Idea

Use Semantic-Web technologies to represent domain concepts and data through a reliable and standardized format.

Main potential advantages of such an approach regard:

- Extensibility: compatibility with data model of forensic tools input and output;
- Flexibility: through custom ontologies definition;
- Information Integration: the RDF data model simplify integration of data coming from multiple sources;

Idea

Use Semantic-Web technologies to represent domain concepts and data through a reliable and standardized format.

Main potential advantages of such an approach regard:

- Extensibility: compatibility with data model of forensic tools input and output;
- Flexibility: through custom ontologies definition;
- Information Integration: the RDF data model simplify integration of data coming from multiple sources;
- Inference: the RDFS/OWL combination and the use of a reasoner can infer class membership and typing information;

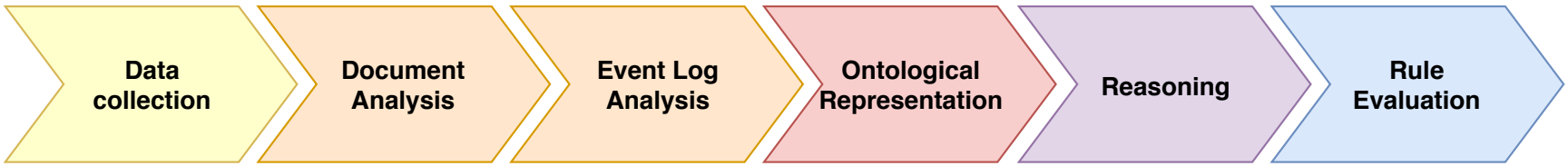
Idea

Use Semantic-Web technologies to represent domain concepts and data through a reliable and standardized format.

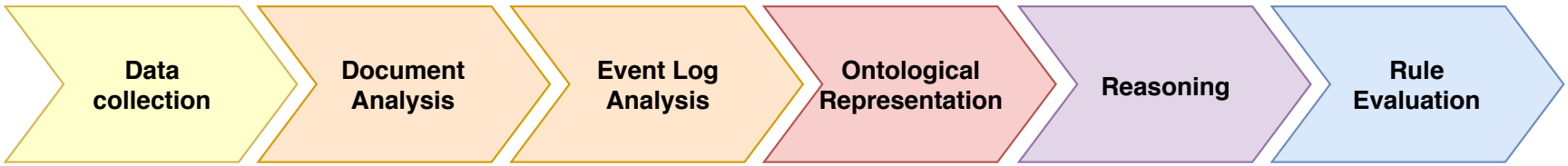
Main potential advantages of such an approach regard:

- Extensibility: compatibility with data model of forensic tools input and output;
- Flexibility: through custom ontologies definition;
- Information Integration: the RDF data model simplify integration of data coming from multiple sources;
- Inference: the RDFS/OWL combination and the use of a reasoner can infer class membership and typing information;
- Search: reasoning engine and the semantic mark-up can enhance queries results.

Methodology



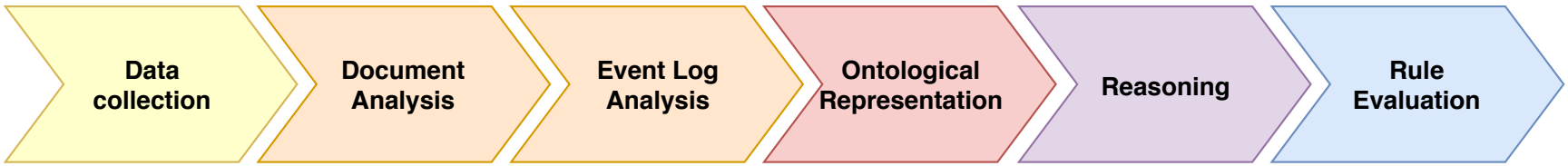
Methodology



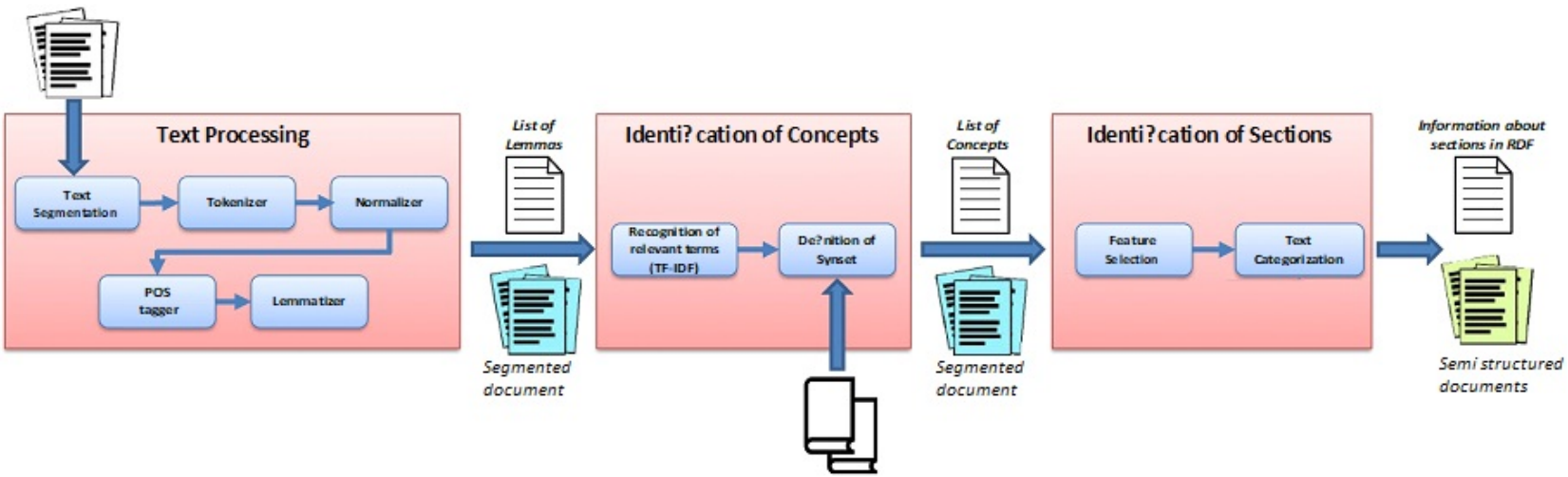
Data Collection

- Involves all the acquisition operations and aims at generating inputs for next phases.
- During this phase preprocessing and data reduction may be applied too, in order to reduce the amount of data managed by next phases.

Methodology

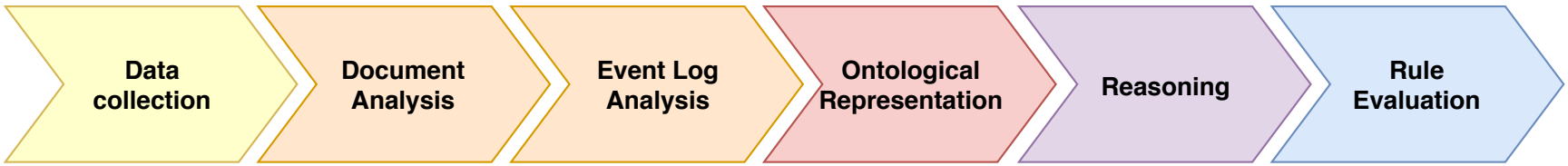


Document Analysis



In order to extract relevant terms from a text, the proposed approach hybridizes linguistic and statistical techniques.

Methodology

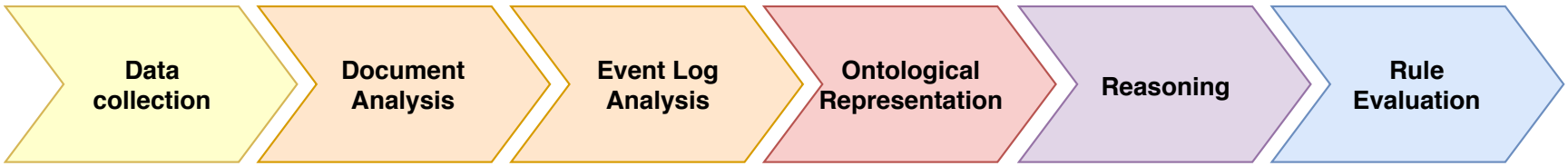


Event Log Analysis

- Attach the results of NLP process (semantic metadata) to the log file chunks, and
- point them to concepts in an ontology.

Information can be exported as a text file annotated with links to the ontology, useful for further ontology population or for indexing to provide semantic search

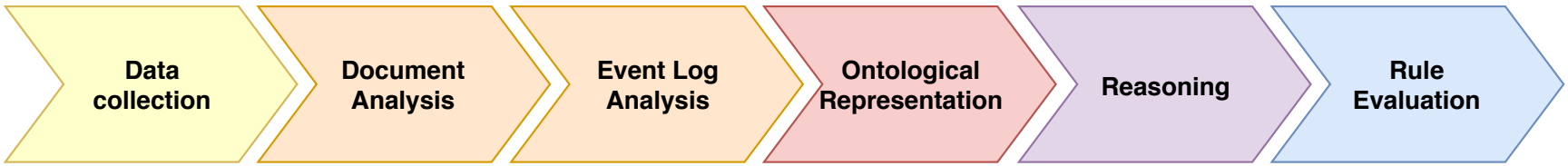
Methodology



Ontological Representation

- Transform the acquired data into a set of triplet constituting the RDF data model.
- Ontologies used in this step can be created ad-hoc or even fetched from shared repositories.
- Outputs can be stored using different formats, such as RDF/XML or RDF/OWL.

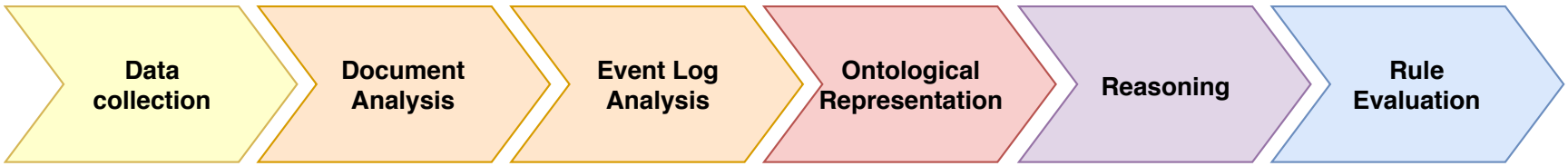
Methodology



Reasoning

- An OWL-based reasoner can infer different types of new axioms
  - classes
  - subclasses
  - properties
- The reasoner can dynamically classify and correlate instances
  - thanks to subclass hierarchy, property relations or property restriction

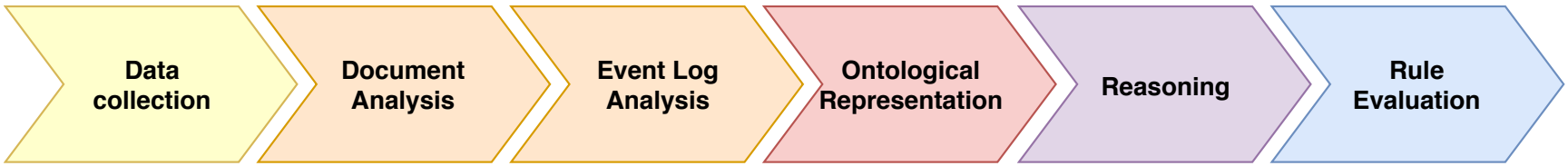
Methodology



Rule evaluation

- SWRL Rules can assert additional axioms that cannot be inferred through OWL.
- A Rule Engine can:
  - evaluate SWRL Rules
  - infer new axioms
  - translate inferred axioms into RDF data model to permit ontology integration

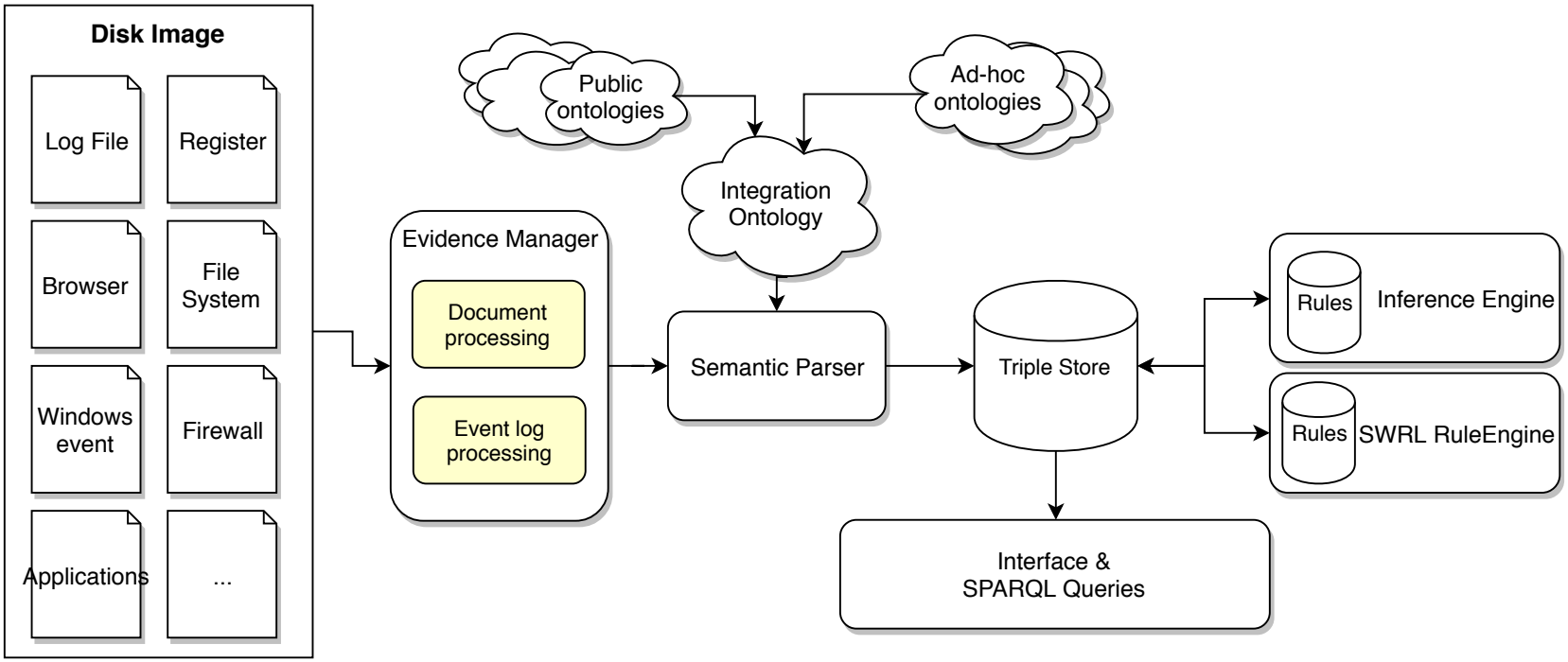
Methodology



Query

- It is possible to query endpoint hosting for the sets of triplets, usually a triple-store
  - by using SPARQL language
  - to search for asserted or inferred axioms.

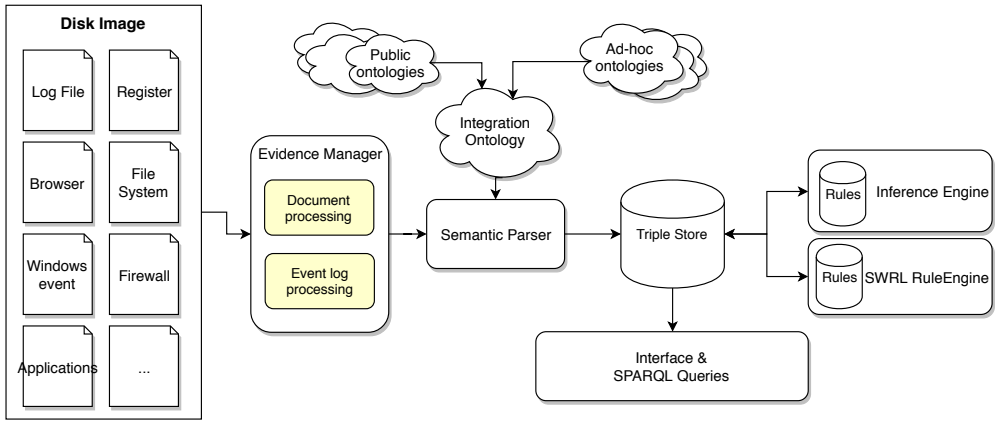
System Architecture



Our overall system consists of an ontology and five modules: Evidences Manager, Semantic Parser, Inference Engine, SWRL Rule Engine and a Query and Visualization module

System Architecture

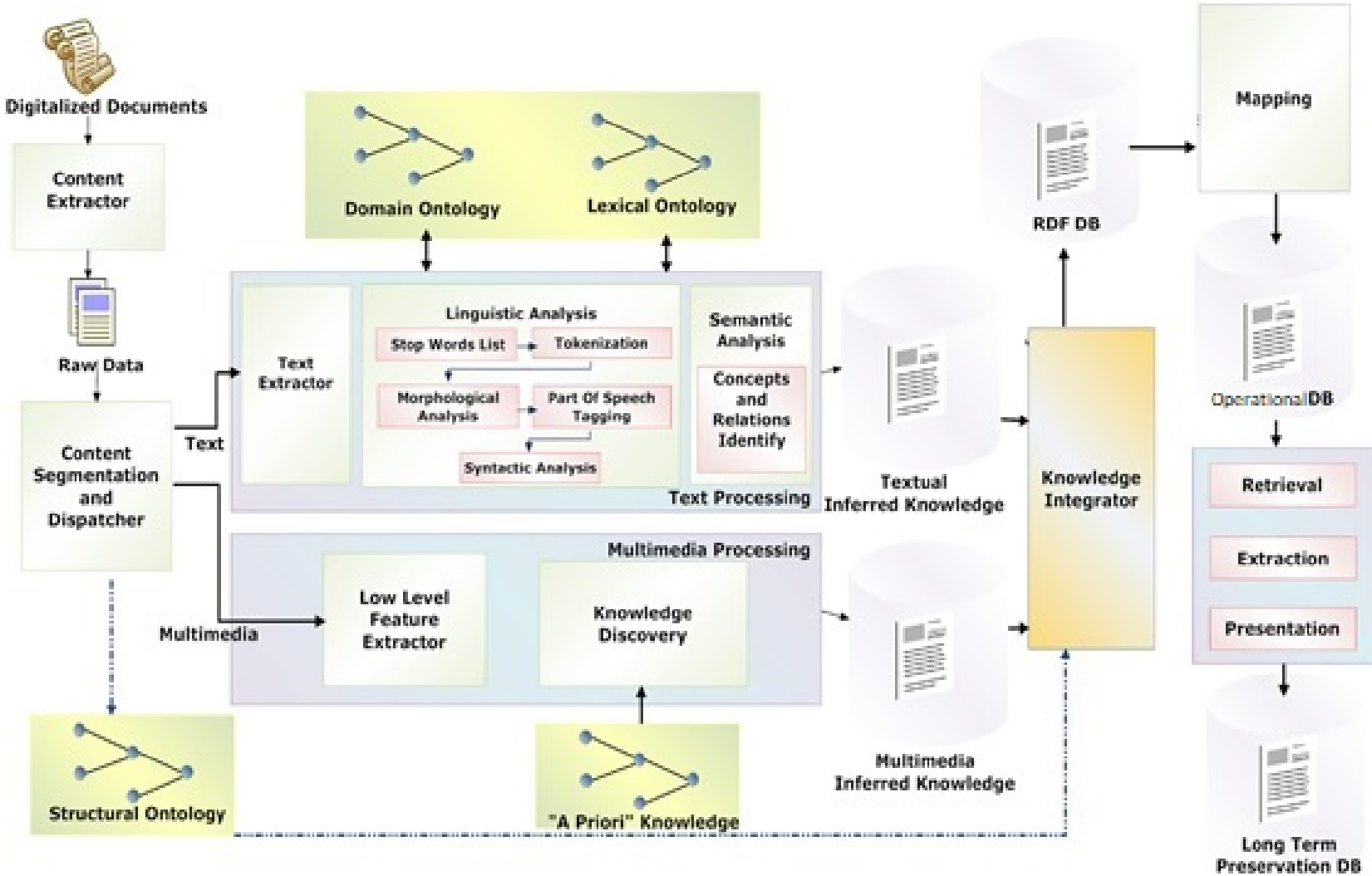
Evidence Manager



- It identifies the type of given source and collect its footprints.
- The extracted knowledge consists of a set of attributes
  - Many of them are already structured and ready to be used
  - non-structured fields, like description of the event, require further processing based on regular expression or NLP techniques.
- The output of this module is a file containing all the footprints retrieved from the source device.

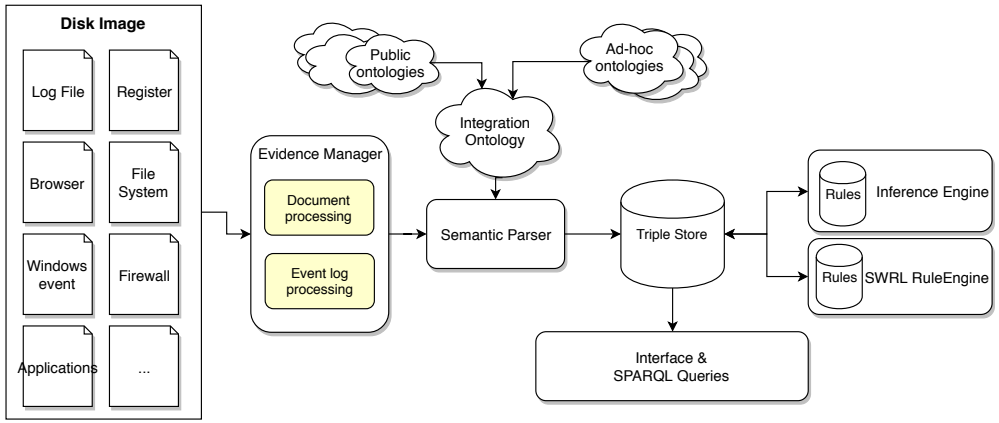
System Architecture

Document Analysis



System Architecture

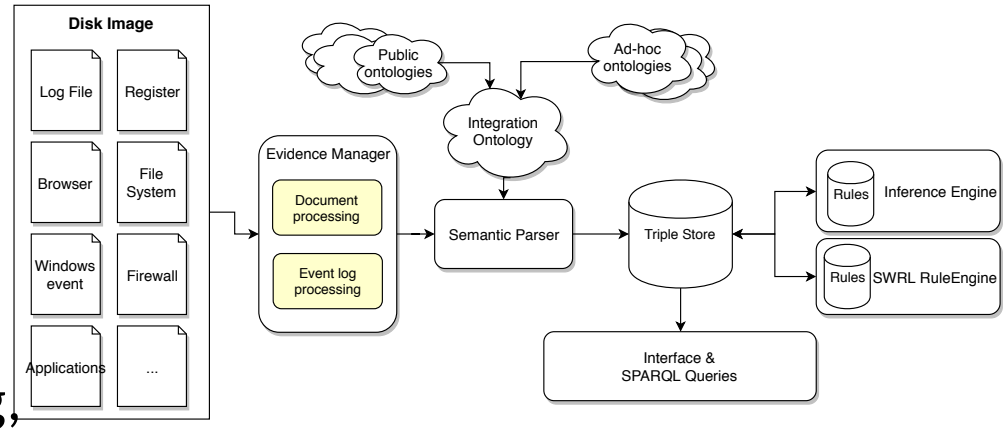
Semantic Parser



- It instantiates the ontology, combined from public domain ones and custom one.
- Ontology population is made by:
  - creating an instance for each footprint item of acquired data
  - linking them each other according to formal object properties defined in the ontology schema.
- It generates a RDF representation of items collected in the previous step.
- The triplets are loaded into a triple store.

System Architecture

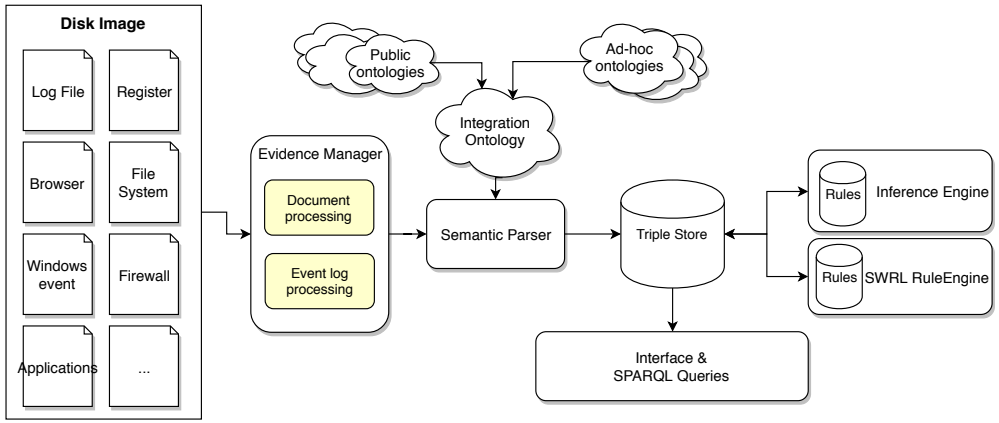
Inference Engine



- It performs automated reasoning, according to the ontology definition.
- It aims to enrich the knowledge base with new inferred facts.
- Examples of inferences:
  - file type from its extension
  - author of an action from user information related to logged account.

System Architecture

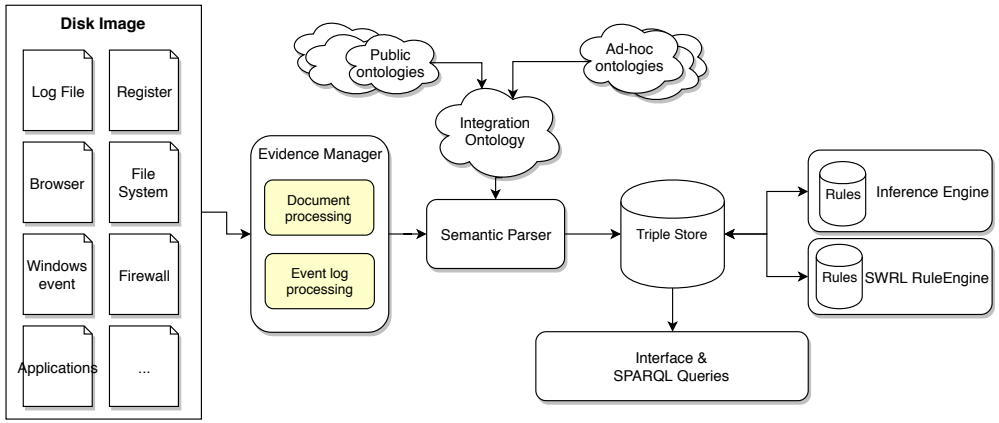
SWRL Rule Engine



- It adopts SWRL rules in order to extend the expressivity of OWL.
  - establish relationships among individuals belonging to different ontologies but representing similar concepts.
- This module can be called
  - before the Inference Engine, make it able to process the axioms that can only be inferred by the SWRL rules
  - twice, one before the Inference Engine and the other after it.

System Architecture

SPARQL Queries



- It is responsible for accepting SPARQL queries from the user.
- It evaluates them
  - using a SPARQL query engine
  - against the set of RDF triples
    - » asserted during the semantic parsing of the source data
    - » inferred by the Inference or the SWRL Rule Engine

Document Analysis Evaluation

- To set up the experimentation, a disk image of the virtual machine containing the dataset is prepared in order to be fed to the Evidence Manager Module
- Then, before being used as input of the Semantic Parser Module, all the evidences were pruned of all known files, such as operating system files with known hash.
- The objective is to evaluate the system correctness during the Document Analysis and Event log Analysis in automatically discovering relevant concepts within a forensic and in particular:

  - User Personal Data
  - System Events (Logon, Shutdown, Application run, etc.).
  - Timeline of Significant Events (e.g., Browser chronology, instant messaging activities, File creation/editing/deletion, etc.).

Document Analysis Evaluation

Relevant concepts discovery procedures exploit a domain ontology built from scratch from the dataset, with the help of domain experts. The comparison between the proposed system semantic indexing output and the ground truth relies on the well-known *Recall* and *Precision* evaluated on the test set (24,000 documents).
The proposed method achieves an average recall value of 94.4% in finding a single concept for recall with respect to an average precision value of 83.7%, a 79.5% recall rate with a 86.8% precision rate in finding ten different concepts and, finally, an average recall of 74.4% with a precision of 97.3% in finding 20 concepts.

Log Analysis Evaluation

- To measure ad-hoc information retrieval effectiveness in the standard way, we used a test collection consisting of three things:
  1. A log file events collection
  2. A test suite of information needs, expressible as queries
  3. A set of relevance judgements, standardly a binary assessment.
- For each event in the test collection is given a binary classification as either relevant or non-relevant. This decision is referred to as the *gold standard* judgement of relevance.

Log Analysis Evaluation

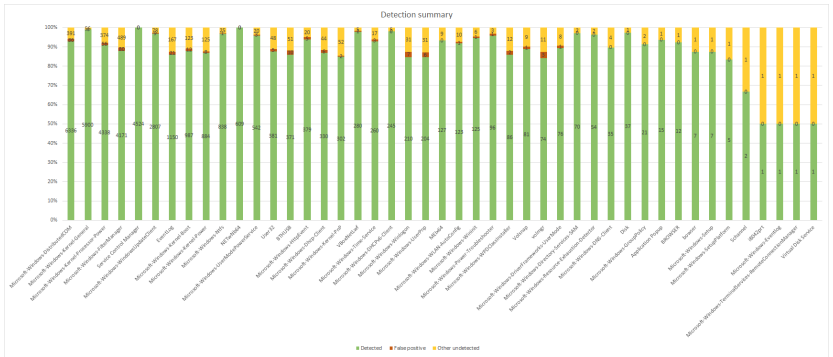
Events are classified into four main categories (Critical, Error, Warning, Information) and for each category we further annotated each event accordingly to the event description. Table shows the number of events of each category for the considered dataset.

Category	**N. of events**
Critical	42
Error	7344
Warning	1193
Information	31102
Total	**39682**

Table: Number of events for each category of considered dataset.

Log Analysis Evaluation

Figure shows that the adopted methodology can detect a large fraction of relevant events in the test dataset.



Detection rate is greater than 85% for those categories containing a considerable number of events. There are also some false positives, which are inevitable in any unsupervised technique, but the rate is never greater than 3% of each considered category.

Semantic pipeline Evaluation

- For this experimental campaign, we generated three datasets
  - populated during the Collection phase from different disk image
- We monitor ontological representation and the enhancement (reasoning and rule evaluation) phases
  - counting for generated and deduced triples.
- we evaluate the correlations by a set of SPARQL query.

Semantic pipeline Evaluation

- For this experimental campaign, we generated three datasets

Table: Volumes of processed data through the reconstruction and analysis processes

Criterion	Dataset No1	Dataset No2	Dataset No3
Extracted entries	6798	4263	15983
Generated triples	8786	10475	32498
Deduced triples	18	21	31
Correlations	1231	535	14982

Semantic pipeline Evaluation

- For this experimental campaign, we generated three datasets

Table: Execution times of methodology phases

Steps	Dataset No1	Dataset No2	Dataset No3
Collection	1.07s	1.43s	1.67s
Representation	24.37s	18.3s	167.45s
Reasoning	0.29s	0.247s	0.314s
Analysis	776.3s	241.3s	12314.7s

Conclusions

- In this work we propose a reusable methodology based on semantic representation, integration and correlation of digital evidence and an architecture that implements it .
- The use of an ontology allows for the representation of knowledge with a unified model and for simplifying the building of analysis processes

Publications

[1] *Improving results of forensics analysis by semantic-based suggestion system.*

[2] *An application of semantic techniques for forensic analysis.*

[3] *An advanced methodology to analyze data stored on mobile devices.*

[4] *Semantic Analysis of Social Data Streams.*

[5] *Using semantic tools to represent data extracted from mobile devices.*

[6] *Detect and correlate information system events through verbose logging messages analysis.*