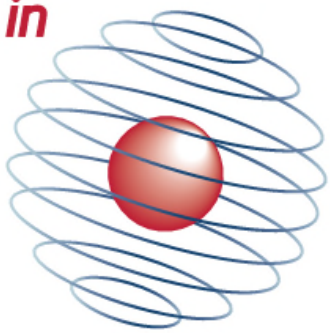
CENTRE *for* DOCTORAL TRAINING *in*
CYBER SECURITY



CDT Technical Paper

06/14

Online Banking Malware Ontology

Rodrigo Carvalho

Online Banking Malware Ontology

Rodrigo Carvalho, *University of Oxford*

Abstract—Due to the ever increasing popularity of the Internet, institutions are migrating their services to the digital realm. Banks are among the most representative examples: in order to better meet their clients' requirements, but also to reduce operational costs, online banking platforms were created and their use stimulated. However, the users' mass adoption to this novel technology without proper awareness campaigns resulted in a large increase of online banking fraud occurrence rates. This poses great challenges to Law Enforcement Agencies dedicated to cybercrime investigation: in addition to personnel skills training, there is an urgent need for new approaches correlating the horizontally sparse and concealed evidence resulted from such offence. As semantic technologies enable the more intelligent use of computer resources regarding data from a specific domain, this paper proposes the creation of an online banking malware investigation ontology.

Index Terms—Ontology, cybercrime, malware, forensics, investigation

1 Introduction

CYBERCRIME tackling is a growing challenge for Law Enforcement Agencies all over the world, and stems from the intrinsic characteristics of the environment where it happens (i.e the WEB). According to the Interpol site, "More and more criminals are exploiting the speed, convenience and anonymity of the Internet to commit a diverse range of criminal activities that know no borders, either physical or virtual." [1]

Due to the inconsistent occurrence detection, it is very difficult to precisely measure global cybercrime activity, such as phishing scam. Even so, some organizations monitor the evolution of attacks reported by specific sources along the year. Among other interesting findings in its 2014 first quarter report [2], the Anti-Phishing Working Group (APWG), an international coalition that brings together several relevant institutions affected by cybercrime, stated that:

- The number of phishing sites leaped by 10.7 percent over the fourth quarter of 2013;
- The number of phishing attacks observed in Q1 was 125,215. That is the second highest number of sites detected in a first quarter, eclipsed only by the 164,032 seen in the first quarter of 2012;
- Payment Services continued to be the most targeted industry sector.

Probably, one of the reasons for such alarming statistics is the investigation complexity. Pieces of evidence from a single offence might be spread in servers and desktops across different countries. Worse still, an United Nations report on cybercrime affirms that "widespread reliance on slow-moving traditional mechanisms such as mutual legal assistance, the emergence of country cooperation clusters, and a lack of clarity on permissible direct law enforcement access to extraterritorial data present challenges to an effective global response". [3]. Even if all countries agreed at once in sharing cybercrime information, there would still be a vocabulary barrier to transpose: different technical and legal terms used by each one would make data integration a non trivial task.

In addition, due to its "novel" investigation skills requirements, there are not enough trained law enforcement officers to tackle cybercrime accordingly, what certainly increases the work backlog. Moreover, the capable ones might get overloaded with the amount of data

• *R. Carvalho is a doctoral candidate at the Centre for Doctoral Training in Cyber Security, University of Oxford.*
E-mail: rodrigo.carvalho@cs.ox.ac.uk

to be analysed, hampering relationships and patterns discovery.

Therefore, it is imperative to use the current computer resources in a more intelligent way, in order to better exploit all the cybercrime evidence stored in either open and closed sources. An important first step would be the "construction of a common language and a set of basic concepts about which the security community can develop a shared understanding...a common language and agreed-upon experimental protocols will facilitate the testing of hypotheses and validation of concepts." [4]

1.1 Online Banking Malware Investigation in Brazil

Bank customers are some of the most common targets of phishing scam attacks. This is particularly true in Brazil, where the fast development of the online banking sector was not accompanied by adequate public awareness campaigns regarding its risks and necessary precautions. As a consequence, a whole Online Banking Malware (OBM) cybercrime ecosystem promptly emerged.

Then, in response to the steep increase of OBM crimes against the government banks, the Brazilian Federal Police (DPF) created the SRCC in 2003, a specialized cyber unit that also supports other cybercrime investigations such as online child abuse and illegal medicine trade. In addition, SRCC is responsible for fostering capacity building among police officers, acquiring specialized software and hardware and also developing bespoke cybercrime investigation tools.

Due to the high number of multiple open investigations that could relate to the same criminal organization, DPF started "Projeto Tentaculos" in 2009. Its main objective was to look for indicators that would allow merging different online banking fraud cases. Its outcomes enabled investigators to go after high hierarchical criminals and dismantle more complex organizations.

Despite its great success, it is still very difficult to identify the malware developers. There are two main reasons for that:

1) Most of the data that feeds "Projeto Tentaculos" comes from the banks' databases, only containing information such as account numbers and holders, date and transferred values. It provides no further clues about the attacks performed;
2) Due to the overwhelming number of fraudulent transactions sent by the bank on a regular basis, the agents focus on investigating the money transfers, which they are more used to, and already count on existing specialized tools and methodology.

Because the same malware is normally sold to and used by multiple thieves, which outnumber the amount of developers, arresting the latter could have a great cascade effect in reducing fraudulent transactions. As a secondary but not less important consequence, more effective OBM investigations could dissuade developers into malware programming.

However, malware evidence is only produced at a different stage, not necessarily connected with "Projeto Tentaculos". Each computer forensics analyst examines all the devices described on a specific search warrant, which sometimes lists multiple large-capacity storage ones. The evidence found in each equipment is then materialised in an Official Forensics Analysis report, which has two main objectives: to substantiate the related criminal case's trial in court, and also to feed the police database with relevant information, further refining current investigations.

Computer forensics analysts's empiric knowledge suggests that while most thieves take few precautions about the evidence left in their computers, malware developers are much more careful, and take account on their technology expertise to hide any traces leading to them.

Nevertheless, the fact that many criminal organizations use malware from a unique developer might increase the chances of finding relevant clues. Thus, it is imperative to collect, analyse and correlate every little piece of information found in devices belonging to the OBM ecosystem. Semantic technologies could play a crucial role in this task, as investigations would be greatly enhanced if the computer could help reasoning about such horizontally sparse data.

1.2 Ontologies

One possibility towards the more intelligent use of computer resources is to provide it with semantic capabilities, and the first step is to create a knowledge base, in which data is stored together with its meaning. This is achieved by developing an ontology, or "a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application." [5]

By establishing a common vocabulary about a domain's concepts and relationships which is understandable by both human and computer agents, an ontology enables, among other features [6]:

- **To share a common understanding of the structure of the information among people and software agents:** once agreed and implemented, data from different sources (e.g. law enforcement agencies from distinct countries) would become compatible, even if not shared yet; additionally, this could make human agents training and software systems developing more homogeneous and integrated;
- **To enable reuse of domain knowledge:** there are many common concepts relevant to distinct domains (e.g. the time concept is important for both cybercrime investigations and chemistry experiments). If such concepts are well implemented and maintained in a specific ontology, researchers from different areas would be able to extend them, avoiding unnecessary rework;
- **To make domain assumptions explicit:** it is easier to notice, understand and change domain assumptions when they are integrated with the data, resembling the way the brain works. On the opposite, knowledge maintenance gets more complex if "raw" data is separated from their meaning, embedded in software source code constructs.

This way, the creation of an ontology would enable, in addition to automated computer reasoning, data input reliability, seamless information sharing, easy knowledge maintenance and homogeneous training across different actors.

1.3 Objective and limitations

This paper will suggest the initial version of an ontology whose main objective is to map OBM criminal organizations and possibly identify the malware developers. Its concepts, properties and relationships will be based on common tasks performed by computer forensics analysts and police investigators.

Due to time and objectivity constraints, it is not a goal of this paper to provide a precisely implemented and tested ontology, nor create a working knowledge base. Instead, it will focus on explaining the reasons why semantic technologies could enhance cybercrime tackling by giving practical examples, and also discussing other ontologies's concepts extension and the need to create novel, OBM-tailored ones.

The forensic analysis and investigations conducted by the Brazilian Federal Police will guide the OBM ontology creation, as they provide the necessary domain knowledge and already implemented guidelines. However, this ontology will be implemented as generic as possible, in accordance to its fundamental aim of fostering distinct stakeholders adoption.

2 Literature Review

In recent years, many researchers acknowledged the benefits of ontologies, which may have caused their migration from the realm of Artificial-Intelligence laboratories to the desktops of domain experts [6]. A notable example is the Semantics Web [7], an attempt to better integrate data from disparate sources on the Internet so it can be shared and reused more rationally. Additionally, fields like biology and medicine are also exploiting its potential considerably [8] [9].

Different tools and methodologies have been proposed to support such initiatives. The following sections will cite some of them, as well as discuss topic-related papers.

2.1 Articles

Among the attempts to employ semantic technologies to the criminal realm, the ”Ontology-based decision support system for crime investigation processes” was suggested in 2005. According to its authors, such framework would optimize information collection, storage, processing and exchange, in order to better support decisions regarding ” the knowledge of the crime scene investigation tactics and strategies of various types of crimes and their peculiarities, where to look for traces, what investigation plan to make up and what problems to solve.” [10]. Although providing a systematic description of crime investigation workflows and suggesting an ontology representing general crime concepts, it does not include cybercrime and the related digital evidence in its scope.

Then, in 2009, a Cyber Forensics ontology was proposed [11], linking the different subclasses of the concepts ”Law”, ”Crime Case”, ”Criminal”, ”Crime Type” and ”Evidence”, the latter further describing collection procedures. As it presents a top-level approach regarding any crime that could leave digital evidence, it does not delve on the analysis of the evidence content itself, but focus on the medium in which it was found (e.g. a memory stick or a disk image). Notwithstanding the importance of general cyber evidence categorization, the OBM ontology proposes a different aim: reasoning among information related to a specific type of cybercrime.

In 2013, researchers from the Computer Emergency Response Team (CERT) at Carnegie Mellon University published a paper discussing their attempt in creating an ontology dedicated to malware analysis, based on established vocabularies and taxonomies. Their goal is to provide a more scientific approach to malware research, and they hope other experts will adopt such ontology, thus starting to speak the same ”official language”.

Although sharing some similar concepts, the OBM ontology has a broader scope, also taking into account the criminal aspect (e.g. the entities responsible for the malware development and use against online banking fraud victims). In addition, CERT’S ontology unavailability prevented further concepts analysis, leading to the specification of new ones.

The ontology proposed in this paper is unique in the way it merges some of these topics with the investigation needs and evidence analysis performed by a Law Enforcement Agency, enabling a task-driven ontology-developing process.

2.2 Methodologies

The implementation of the complete life cycle of an ontology development process is not a simple task. It involves many concepts inherent to software engineering projects, such as resources management, evaluation and testing, activities scheduling and iterative cycles.

Therefore, analysing the different ontology-engineering methodologies (such as Methontology) in order to define the best fit for the OBM investigation domain is not one of the objectives of this paper. For related information, please refer to the NeOn project report [12] that, besides suggesting a new methodology for building ontology networks, also presents a good evaluation between well-established ones.

This paper will follow the Knowledge-Engineering Methodology steps described by University of Stanford researchers in [6], as it favours explaining ontology-specific concepts and issues to the detriment of describing a complete and formal engineering process, thus enabling a better understanding of such technology. Section 3 describes its guidelines.

After defining the methodology to follow, the next step is to choose between the available tools and languages to effectively start building the ontology.

2.3 Tools

Two of the most complete and advanced ontology-building languages are the Web Ontology Language (OWL) [13] and the Nepomuk Representation Language (NRL) [14]. Although they have common origin (both extend the Resource Description Framework (RDF) [15] and the associated RDF Schema (RDFS) languages) and purpose (both represent and

process knowledge in a machine-interpretable way), they were initially targeted at different domains.

OWL was designed to provide semantic capabilities to the Web, allowing automatic processing and integration of data from distinct sources based on its meaning. It became a World Wide Web Consortium (W3C) recommendation in 2004, which has been maintaining OWL since then. Differently, NRL objective was to provide semantic power to desktop applications, by structuring the context of all personal information stored in someone's computer.

Another important distinction between them is that reasoning in OWL is based on the Open World Assumption (OWA), in which the absence of a statement doesn't mean it is false. Instead, its truth value remains undefined, as there might exist unknown information that could directly affect the assertion. On the opposite, NRL is based on the Closed World Assumption (CWA), meaning that any statement which does not hold a true value is considered false. It matches the "...expectations of the (NRL) users better, as a local desktop is indeed a closed world with a limited, known, processable number of files." [16]

The very nature of police investigations suggests that it is more adequate to reason upon OWA (after all, the current lack of incriminating evidence does not necessarily means a suspect is innocent, as future investigations might confirm its guiltiness). However, the capabilities semantic technologies can bring to cybercrime fighting will be demonstrated using NRL, as specialized ontologies which can greatly support the OBM ontology derived from it.

Finally, the Protege ontology editor was chosen to implement the concepts, properties and relationships of the OBM ontology. It is a popular free software that counts on both active community support and tailored plug-ins which make the creation process easier.

3 Ontology Development 101

This section explains important engineering concepts by describing the "Ontology Development 101" methodology, which will guide the OBM implementation. There are seven mains steps:

1) **Determine the domain and scope of the ontology:** specify the realm it is targeted to, and which aspects will be covered. One conclusion from the NeON report [12] is that "...most of the analysed methodologies propose simple methods for carrying out the ontology specification activity. The methods consist of high level steps that can be summarized as follows: identify purpose, uses and users for the ontology to be developed, and identify the set of requirements the ontology to be developed should fulfill". Likewise, the use of competency questions, an informal list of questions that the knowledge base referencing the ontology should be able to answer [6], is suggested;
2) **Consider reusing existing ontologies:** instead of starting from scratch, check whether an ontology covering similar concepts already exists. The more general the concept (e.g. geolocation), the most likely to find a good candidate. Although extending an ontology could optimize concept specification and future data integration, its level of adherence to other ontologies and maintenance support should be carefully assessed;
3) **Enumerate important terms in the ontology:** whether they represent a property, a relationship or and entity, it is recommended to "...write down a list of all terms we would like either to make statements about or to explain to a user" [6] beforehand;
4) **Define the classes and the class hierarchy:** in other words, decide on the main concepts of the target domain and their subsequent specializations, as in a tree graph containing the root node (e.g. *vehicle*), the intermediate nodes (e.g. the disjoint classes *car*, *motorcycle*, *truck*) and the leaves (e.g., *SUV*, *estate*, *convertible* as subclasses from *car* and, consequently, from *vehicle*);
5) **Define the properties of classes (slots):** these are the internal structure of the

concepts, and can be either object properties, stating the possible relationships with other classes (e.g. class *car* relates to class *company* by the *maker* property), and data properties (e.g. class *car* has *color* and *horsepower* properties). It is important to notice that there is not a right way to model an ontology: instead of belonging to the *car* class, *horsepower* could be a property of the *engine* class, which could relate to the former through the object property *isPartOf*. The tasks defined by the competency questions should guide the necessary level of detail for each class in the ontology. Finally, it is worthy mentioning that a class slots are inherited by all of its children (e.g. *SUV*, *estate* and *convertible* would contain the same *colour* data property and the *maker* object property of the class *car*);

6) **Define the facets of the slots:** facets are information describing features of the slots values:
   - **Value type:** defines the type of the data described by the slot. Some examples are *string, number, boolean* and *list of allowed values*;
   - **Cardinality:** specify how many values a slot can have, whether single or multiple. In addition, it can be further refined by establishing minimum and maximum allowed values;
   - **Domain/Range:** basically, they state the classes linked by a specific object property. In the previous example, the property *isPartOf* has the class *engine* as its domain, and the class *car* as its range, so the relationship makes sense: *engine* isPartOf *car*.
7) **Create instances:** this means merging raw data with the ontology, establishing the knowledge base. For instance, defining the instance *Cooper* as a *car* (so it is also a *vehicle* but not a *truck*); the instance *Mini* as a *company* (making it possible to distinguish between *cars* from different *makers*); *red* as a *colour* with range either *string* or *list of allowed values*; and *134* as an integer value describing *horsepower*, with single cardinality (an engine has only one *horsepower* value; a *car*, however, might have two *colours*).

4 Building the OBM Ontology

In order to properly apply the methodology described in Section 3, there are some realm-specific notions and guidelines that might help addressing the requirements of each step. The OBM ontology will be defined as follows:

4.1 Ontology domain and scope

The domain of the proposed ontology is the digital data stored in devices seized by DPF during OBM operations, and it will cover concepts and relationships routinely applied to forensic and investigation procedures. At the first moment, this ontology will be maintained and improved by computer forensics analysts, who together with police investigators will use it for:

1) Map different criminal organizations and identify the malware developers, by uncovering relationships between a great quantity of supposedly unrelated evidence;
2) Enable future data integration between different police forces and improve OBM implications discussing among inexperienced individuals, by providing a standardized way of collecting, storing and representing its information;
3) Speed up the analysis process: as the ontology exposes the peculiarities from each malware variety (e.g. the names given to text files containing victims' bank details), forensic analysts could consult it for clues on evidence not yet found.

Finally, competency questions might help narrowing down the ontology scope. Some examples are:

- Is there any relationship between different criminal organizations using the same malware?
- Who developed this malware?
- Are specific malware varieties being mostly used in particular regions?

- What are the most effective phishing scam types?
- Is the recipient email address contained in the malware executable the same in all of its versions?
- Who is the owner of this email address?
- Which suspects know each other?

4.2 Existing related ontologies

Despite semantic technologies' recent incorporation in diverse areas of knowledge, a multitude of ontologies have been developed. Ranging from professional, well-maintained and discussed implementations to single person initiatives, there are good chances that an ontology containing similar concepts to the proposed one already exists.

Although Google-like search engines were created to look for specific keywords across different ontologies terms, specification and instances, most of the results returned by some of them are outdated or unavailable. Therefore, finding a good related ontology might be a matter of browsing through projects homepages and forums. The VocabularyMarket wiki [17], maintained by W3C, is a recommended starting point for such task.

In addition to the ones mentioned in Subsection 2.1, the proposed OBM ontology extends many concepts from the OSCAF ontologies project, which derived from "...the contributions to a number of large-scale efforts, starting with the European project NEPOMUK (2006-2008), continuing with their adoption (and extension) by the KDE community (2009-2013), and their further extension by a second European project Digital.Me" [18].

It applies the NRL in order to provide high-level knowledge representation for all user-related data stored in the so-called Semantic Desktop, like files, contacts and messages (through the Nepomuk File, Contact and Message Ontologies respectively - NFO, NCO and NMO). Both its scope alignment (after all, most crime evidences are contained in similar data structures) and its long term support and development history make the OSCAF group of ontologies an excellent foundation for the OBM ontology creation process.

Moreover, any schema for crime investigation must necessarily represent the agent committing the offence. To achieve that, an important concept from the Friend of a Friend (FOAF) ontology will be extended: it defines an agent as a "thing" (a person, group or software) that do things. This is particularly useful for the proposed ontology, as it also considers malware an entity due to its capability of performing different tasks like creating text files or sending emails. Further details will be discussed in Subsection 4.4.1.

Whilst implementing similar concepts to the NCO, like the *name*, *age* and *gender*, FOAF focuses on the web activity of the instance, describing its related home pages and mailboxes. In fact, as the "local activity" of the user is merged with its "online activity" (for instance, cloud storage of personal files) the NCO also incorporates web-related concepts such as *instant messaging IDs*. In addition, it also provides a more generic approach to the contact's concept definition, as " every piece of data that identifies an entity or provides means to communicate with it." [19], and thus will be the main source for contact-related ontology terms.

4.3 Important terms

By defining a list of important terms to be considered, it is possible to realize whether it will be necessary to define new concepts, which are not covered by the previously consulted ontologies. Table 1 lists some examples from the OBM domain, based in both the author's empiric knowledge and the content of 10 selected Analysis Reports, produced by DPF's Computer Forensics Experts during 2013. These sources of information will also guide the next Sections contents' discussion.

4.4 Class hierarchy and properties

Different strategies can be implemented to define the class hierarchy: top-down (concepts specialization beginning with the most general one), bottom-up (grouping specific classes to form more generic concepts) or a combination of both. The paper will implement the latter approach, as it considers the most relevant

malware	file	person
version	size	nickname
downloader	hash	criminal
comments	date	sms
source code	accounts	company
email	passwords	contact
phishing	remote server	c&c
attachment	log	bank
url	strings	file path
investigation	forensic	container
member	instructions	location

TABLE 1: Suggested terms for the OBM Ontology.

identified terms (whether generic or specific) as a good starting point for driving the ontology engineering.

Among the terms listed in Table 1, the fundamental chosen concepts to start building the classes tree are *malware* (one of the main evidences that the offence happened), *person* (mainly the developer and the criminal), *knows* (an important relationship between people that can help mapping the OBM crime ecosystem) and *email* (the primary way of communication, regarding both phishing scams and messages sent from the malware containing victims' account and password details).

The following subsections will propose the OBM ontology class and properties hierarchy that, despite extending some of the OSCAF concepts, also implements new ones, intrinsic to both the forensics and investigation domains. Additionally, related semantic queries that could optimize evidence finding will be suggested.

4.4.1 Entity

Inspired on the "agent concept" from FOAF (a thing that does things), OBM will implement the class *Entity*, containing the concepts Person, Group and Malware, as depicted in Figure 1.

The actual malware is considered an entity due to its capability of taking actions based on the feedback from the environment, which resembles the ones performed by a "real" thief: presenting a bait (phishing scam) to deceive a naive person, writing down the victims' bank details (appending them to text files) and delivering a list containing multiple victims' information to the gang chief (through emails or

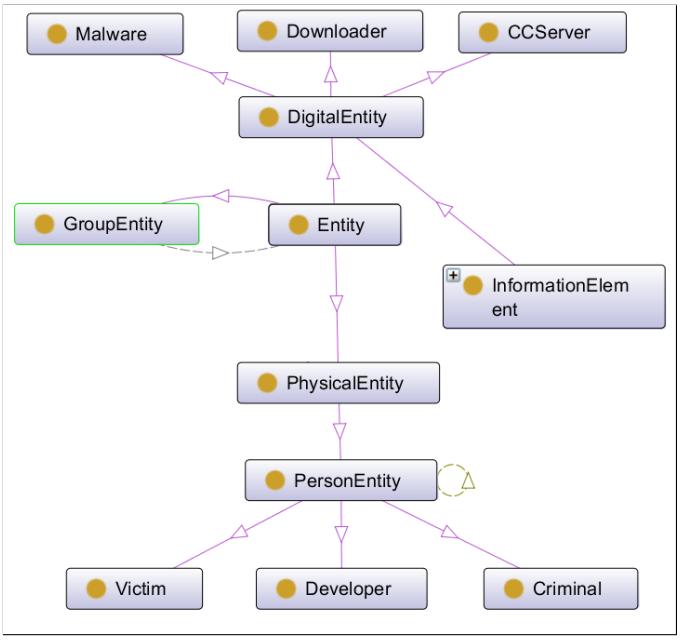


Fig. 1: Entity class and subclasses.

Short Message Service (SMS) messages.)

One might argue that as there are multiple copies of the same version of a malware within the OBM crime ecosystem, this would undermine its bonding with specific organizations. However, their hash codes might be different. For instance, in order to communicate with human entities, each malware would have a different email recipient embedded in its executable.

Therefore, inference rules regarding the hash values of malware found in different seized devices, such as the ones depicted in Figure 4, could allow the computer to reason if their owners are related to the same organization. The same affirmation can not be made to "static" evidence, such as the pictures used in a forged bank homepage. Because their content is not normally customized across different criminal organizations, they would solely indicate that a crime offence happened in that device.

However, it is still necessary to distinguish between physical and digital entities, as there are intrinsically different ways to describe and relate them: while someone might have a *postal address* and *know* (the symmetric and reflexive relationship denoted in yellow in Figure 1) another *PersonEntity*, a piece of software can be identified by its *hash*, and might contain clues

about its developer in the executable file (such as the project folder from which the source code was compiled). This is the reason why *Digital Entity* is also considered a subclass of *InformationElement*, further detailed in Subsection 4.4.3.

Finally, the leaves of the *Entity* class tree reflect the main types of *DigitalEntity* and *PersonEntity* involved in an OBM Investigation. Their data properties (e.g. nicknames and connected URLs) might also optimize relationship discovery.

4.4.2 ContactMedium

In addition to phone calls, which are excluded from the scope of this paper, OBM groups members contact each other through email, SMS and IM messages. Therefore, the *ContactMedium* class was extended from NCO, as it implements the *PhoneNumber*, *PostalAddress*, *EmailAddress* and *IMAccount* subclasses.

Moreover, as the OBM ontology considers malware as an *Entity*, the *TCPIPAddress* class was created. Its main purpose is to represent communication between different kinds of *DigitalEntities*. For instance, a downloader requesting a malware from a specific server: a rule could account on the level of similarity (perharps by the number of identical function calls) among malware obtained from the same server, but by different downloaders. A high level of similarity could then indicate that the developer might be maintaining the server and supporting a specific group of criminal organizations.

Figure 2 shows the *hasContactMedium* object property between *Entity* and *ContactMedium*. Also, that any *Message* can be *from*, have a *recipient* or *reply* to an *Entity*'s *PhoneNumber* (in the case of SMS), *EmailAddress*, *PostalAddress* (domain restricted to *PhysicalEntities*), *TCPIPAddress* (domain restricted to *DigitalEntities*) or *IMAccount*. The latter has a crucial role in establishing the *know* object property among *PhyisicalEntity* instances, and will be better discussed in Subsection 4.4.3.

Other relevant information such as *sent date* are described in the data properties of the *Message* class, as they do not relate directly to an *Entity*.

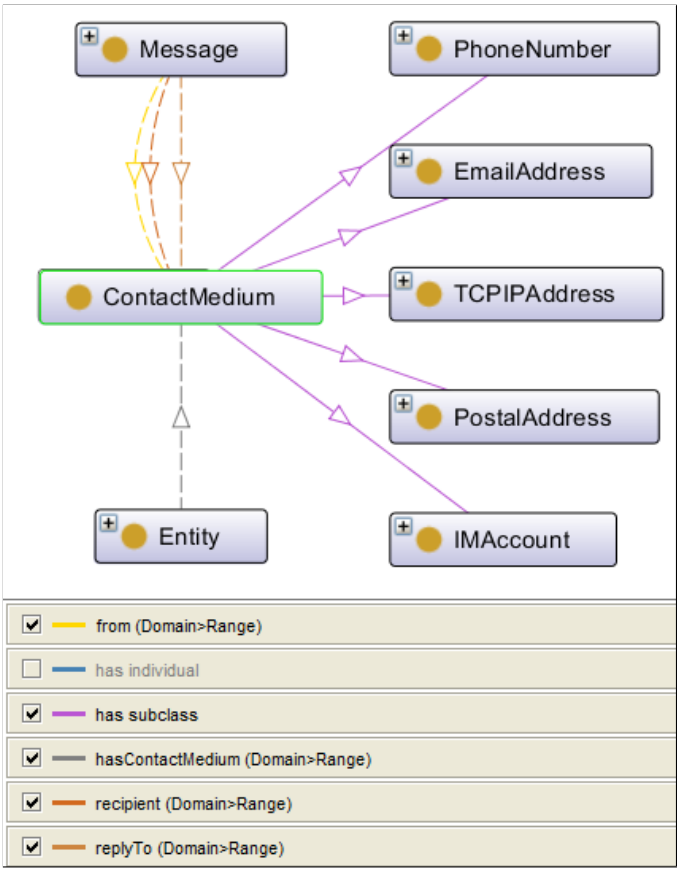


Fig. 2: ContactMedium class and subclasses.

4.4.3 Information Element

The *Message* class mentioned in Subsection 4.4.2 is a child from the *InformationElement* class, along with its siblings *Contact, ContactList, DigitalEntity and Document*. The OSCAF ontologies makes a clear yet very appropriate distinction between an *InformationElement* (describing content-specific information) and a *DataObject* (representing the "physical" container). They are connected through the *isPartOf* relationship: a *File is*(the physical)*PartOf* a *TextDocument*, for instance.

This approach provides the necessary level of flexibility for describing digital evidence commonly found in a seized storage device. The NIE specification [20] gives the example of the mailbox, an *InformationElement* subclass that can be represented by either a local *FileDataObject* (e.g. "inbox.pst") or a *RemoteHostAddress* (in the case of the IMAP protocol). Although having different representations, the interpretation (mailbox) remains the same.

The InformationElement classes extended to

the OBM ontology, depicted in Figure 3, are:

Message

Represents the main messages exchanged in the OBM crime ecosystem:

1) Regular email and IM between criminals: in addition to sender and recipient usernames, cited nicknames, locations or any other identifiable *contact* could be linked using the *relatedTo* property. This could allow the computer to analyse all messages exchanged by two entities, looking for their most probable nicknames, for instance;
2) Regular email and SMS between malware and criminals, containing victims' bank details in text format;
3) Phishing emails between spam senders and victims, which contains a link to an evil *URL*. Similarity measures could be established regarding the IP addresses pointed by them;
4) Control signals between different types of malware, informing the remote C&C server that a victim has just connected to the online banking site; or a downloader requesting a malware from a remote server.

Document

Different documents related to OBM investigation are encountered during the forensic analysis, most of them text files from the criminal seized device: instructions for malware usage, email addresses listings, intercepted information (name, account number, passwords, among others). Often they include comments with specific words also be found in different documents. Interpreting such information using the *keyword* data property could help relating a great number of documents towards its *producer* identification.

Similar properties, such as *definesClass*, *definesFunction* and *definesProgrammingLanguage* could be also related against different *SourceCodes* to help confirming a supposed unique origin.

Contact

The *Contact* class extended from OSCAF ontologies is broader than a simple person representation within a IM software, for instance. It encompasses every piece of information that can help identifying an entity. Some examples are *nicknames* found in text files and malware *versionNumbers*. This approach is extremely useful for OBM investigation, as it allows collecting and reasoning upon little, atomic information dispersed over different cases.

The *Contact* class would allow, for instance, that an unknown *nickname* found in a text file is inserted into the knowledge base. Because that file's *DataObject* is linked to the owner (*PersonContact*) of the seized device (*DataSource*), this could automatically suggest a weak yet possible relationship with the person referred by that *nickname*. This hypothesis could be latter confirmed or refuted with further added information.

Finally, there are distinct data properties for *PersonContact* and *MalwareContact*. The former contains *fullName* and *birthDate*, and the latter *versionNumber* and *targetBank*, for instance. It is worthy noticing that these are content-specific properties directly relating to an entity. More generic properties that describe metadata (such as *modifiedDate* and *hashValue*) refer to the *DataObject* representing them. More details will be given in Subsection 4.4.4.

ContactList

Being a subclass of *InformationElement*, a *ContactList* would have the *PersonContact* representing the owner of the seized device as its *creator*. In addition, it holds the object property *containsContact*,whose range is *ContactListDataObject*. Each one of these would be interpreted as a new *Contact* in the knowledge base, initially linked to the list *creator*. Then, as more *ContactList* information is collected from different sources, the chances of automatically mapping the recently added *PersonContacts* to the previosly stored *ContactListDataObjects* increase.

DigitalEntity

Despite being able to do things, a DigitalEntity is still a sequence of bytes, more precisely a

software. Therefore, it is also a subclass of *InformationElement*, inheriting object properties (e.g. *isStoredAs*, linking it with the corresponding *FileDataObject*), and data properties such as the previously mentioned *keyword* (useful to represent identifiable strings extracted from the executable) and also *contributor*, describing any *Contact* somewhat related to its content production. Any *emailAddress* or *phoneNumber* found in its executable would not be *keywords* from an *InformationElement*, but instead the actual *ContactMedium* from the current *DigitalEntity*. Nevertheless, they could still be compared to find relationship between different cases.

Regarding this class, the investigator could assess the following hypothesis: two different pieces of malware with the same *name* and invoking more than 10 identical *defineFunctions*, but with different *emailAddresses* embedded in their code, refer to two different organizations using malware from the same developer. However, if the storage devices containing both malwares were seized in the same town, that could suggest some level of relationship between them.

4.4.4 DataObject

The *DataObject* is the container of an *InformationElement*. As explained in [20], "It represents a native structure the user works with. The usage of the term 'native' is important. It means that a DataObject can be directly mapped to a data structure maintained by a native application. This may be a file, a set of files or a part of a file." The relevant subclasses to the OBM ontology are:

CarvedDataItem

Stores information retrieved from the file system non-allocated space. It is created by the recovery tool, and would only indicate that the content container has been permanently deleted. Although its offset could be easily determined, it does not carry enough investigation relevance to be represented in the ontology.

FileDataObject

Comprises files from allocated disk space, whether local, deleted (to the trash bin), remote or embedded ones. It is the most common container for *InformationElements*, and contains relevant linking-capable properties such as *hashValue*, *dateModified* and *fileSize*.

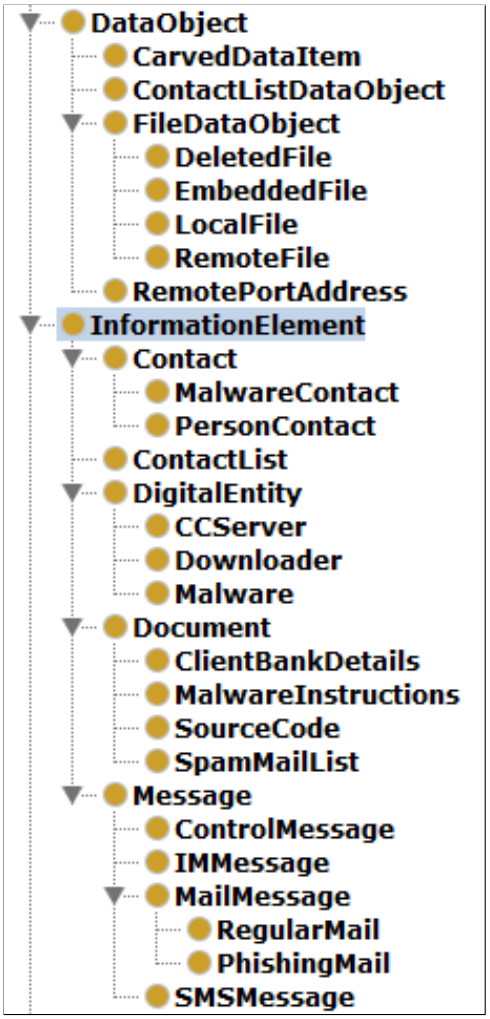


Fig. 3: InformationElement and DataObject subclasses

RemotePortAddress

As stated in the mailbox example, it is a *DataObject* solely represented by the pair IP/Port of the service hosting it. It is necessary to specify the origin of malware downloaded from different servers, which could be related to the evil *URLs* from the *PhishingScam* emails in order to identify common infrastructure shared by different criminal organizations.

ContacListDataObject

Stores each *Contact* within a specific *ContactList*. It needs a specific representation because there

might be multiple *ContacListDataObjects* stored in the same *FileDataObject* (e.g. "contacts.edb"). In addition, it contains relevant investigation metadata such as the date each contact was added.

4.4.5 Other Important Classes

There are some important concepts that, despite seeming to represent data properties at first look, are more useful if defined as classes:

Name

Being defined as a class would allow a specific *name* to be compared against different *contacts* in order to find out, for instance, to which *entity* it belongs to. After all, the same entity can have different names (in this case, *name* serves as the domain of the data property *nickname*), and the same name can be attributed to different entities. This NCO-extended concept also implements other disambiguation classes, such as *BirthDate* and *Gender*;

Hash

Being defined as a class would facilitate grouping files with identical hashes. In addition, as the ontology also provides alternative ways of identifying malware varieties, it would be possible to catalog different *hashValues* from the same malware (e.g. in the case they differ by the embedded *email* recipient). Consequently, this would speed up searching for same-family malwares within a big database.

DataSource

This represents the seized device (e.g. laptop, smartphone) from which information was collected. Instead of simply assigning each an unique item number to be referenced as one of the *DataObject*'s data property, declaring it as the "root physical container" of all digital evidence regarding a specific suspect would facilitate chain of custody's management.

It is imperative to assure the device integrity along its way to the court, so it can be considered valid. Because different people manipulate it, starting at the seizure location, passing by the agency's storage room and finally reaching the forensics lab, object properties such as *MovedBy* and *NextDestination* are essential, and could help identifying insider threats, for instance, by reasoning upon uncommon procedures regarding a specific device.

Finally, other vital concepts for OBM investigation are *Location* and *Date*. These are general concepts whose subclasses and relationships have been comprehensively implemented by specialized ontologies [21] [22]. As an example, they could be extended to find patterns about the period different criminal organizations have been active for and the location of their members, whether *PhysicalEntities* (e.g. the city most of them live) or *DigitalEntities* (e.g. the geolocation of the remote server storing it).

4.5 Facets and Instances

The next step after distinguishing the object properties from the data properties is to define their facets. As detailed in Section 3, three different aspects of the values can be described: value type, cardinality and their domains and ranges. Table 2 lists some OBM properties along with their facets. Column "C" states either single or multiple cardinality, and the "*" symbol distinguishes data properties.

Domain	Property	Range / Type	C
Info.Element	keyword *	string	M
Info.Element	contributor	Contact	M
Info.Element	creator	Contact	1
Entity	hasContact	Contact	M
Contact	hasName	Name	M
Name	nickname *	string	M
Info.Element	isStoredAs	DataObject	M
DataObject	lastModified *	dateTime	1
DataObject	dataSource	DataSource	1
ContactList	containsContact	ContactListObj	M
Info.Element	relatedTo	DataObject	M
Entity	hasContactMedium	EmailAddress	M
EmailAddress	emailAddress *	string	1
Message	from	ContactMedium	1
Message	inReplyTo	Message	M

TABLE 2: Some slots and facets from OBM ontology.

Lastly, instances of the each OBM ontology class should be created and their slots filled up in order to establish the knowledge base. As stated before, its ultimate goal is to enable the computer to answer semantic queries requested by the human user by reasoning upon the inserted data and established axioms.

Figure 5 depicts some of the classes, relationships and properties mentioned in the previous subsections. For the sack of simplicity, some underlying concepts such as *DataObject* were omitted.

5 Discussion and Future Work

Semantic technologies consist in a paradigm shift in the way we deal of data, as it resembles the way the human brain works (yet very primitively) when evaluating the information it has access to. Despite the existence of some well established ontologies, they are very few if compared to the number of candidate areas of knowledge (whether academic, professional or social) which could benefit from semantic technologies.

Therefore, the main result expected from the ontology suggested in this paper is to spark some interest about the great capabilities such techniques could provide to the cybercrime realm. After all, Online Banking Malware is only one among all different investigation scopes that can be derived from the seized digital data domain.

By no means the suggested OBM ontology intends to be a definite implementation, especially because of the intrinsic evolving characteristic of knowledge bases: specific domain assumptions might change over time, and more optimized inference rules can be established by assessing semantic queries outcomes.

Although based on empiric knowledge and official information from current analysis reports, the concepts, relationships and semantic queries proposed by this paper have not been tested against real data. Therefore, it was not possible to assess their practical relevance supporting forensic and investigation tasks.

For instance, whether the seizure location would be a good criteria for defining if two similar pieces of malware, but containing different embedded email address recipients, are members of the same organization. Because there might be no statistics regarding this assumption and many others, they would need to be evaluated in real case scenarios and consider other relevant relationships which could advise towards their validation or refutation.

Another important aspect to reflect upon is the classification of the malware as an entity. Due to the reasons previously presented, it might indeed help mapping different criminal organizations. However, this definition has not been settled among forensics analysts and investigators, at least not yet. If deemed valid, the ontology itself could help disseminating this novel perspective among them.

Finally, the feasibility of automatic information extraction has to be considered, as it is a previous important step towards evidence gathering. For instance, in the case that a big chat history file is found, containing long conversations with different recipients, how would information (e.g. location, nicknames, email addresses and references to victims' bank details) be collected? Whereas this problem might be considered out of the scope, a failure in addressing it would risk the ontology adoption and efficiency, as the amount of necessary work to manually input all this data could discourage some users.

Regarding this and some other previously discussed ideas, some suggestions for future work are:

- Implement a prototype following a well-established ontology-engineering process, which would envisage the necessary formality regarding requirements analysis, software design, implementation and testing;
- Load real data to the knowledge base in order to measure the degree in which the initial goals are met, and also to validate and constantly improve its efficiency and efficacy;
- Define a standards authority for cybercrime ontologies development, recognized and respected by all of its official users. If their adoption rates increase, different stakeholders might need to take part on it (e.g. a partner country's agency);
- Create an official online resource containing the URL to the OBM ontology, its specification and also a forum to centralize suggestions and issue discussions. Its open of restricted accessibility would have to be further debated though;
- Research about Natural Language Process-

ing techniques that would, in addition to automate entity extraction and provide content-based file categorization, create a seamlessly interface between their output and the data format expected by the knowledge base.

6 Conclusion

This paper discussed some current issues related to cybercrime investigation which affect law enforcement agencies, mostly derived from the complexity in finding and relating OBM evidence within the great amount of data sent to forensic analysis. It briefly explained the Online Banking Malware domain within the Brazilian Federal Police and suggested that providing semantic capabilities to the computer could make its investigation more effective and efficient.

An OBM ontology was proposed: its classes, properties and relationships were thoroughly discussed, and sample semantic queries, based on common tasks performed by forensics analysts and investigators, were suggested.

Although the actual creation of the knowledge base was recommended as a future project, this paper expects to spark the interest of semantic technologies to law enforcement officers and decision makers, as it considers that they will cause a paradigm shift to traditional cybercrime investigation, this time in favour of the justice.

Acknowledgments

The author would like to thank Dr. Bernardo Cuenca-Grau for supervising this project; the University of Oxford CDT in Cyber Security team, for their vital assistance; and finally CAPES and the Brazilian Federal Police, for funding and supporting my DPhil programme.

References

[1] "Cybercrime - INTERPOL." [Online]. Available: http://www.interpol.int/Crime-areas/Cybercrime/Cybercrime

[2] APWG, "Phishing activity trends report - 1 st quarter 2014," Tech. Rep., 2014. [Online]. Available: http://docs.apwg.org/reports/apwg_trends_report_q1_2014.pdf

[3] Steven Malby and Robyn Mace, "Comprehensive study on cybercrime," United Nations Office on Drugs and Crime, Tech. Rep., 2013. [Online]. Available: http://www.unodc.org/documents/organized-crime/UNODC_CCPCJ_EG.4_2013/CYBERCRIME_STUDY_210213.pdf

[4] JASON Program Office, "Science of cyber-security," The MITRE Corporation, Tech. Rep., 2010.

[5] Tom Gruber, "Ontology (computer science) - definition in encyclopedia of database systems." [Online]. Available: http://tomgruber.org/writing/ontology-definition-2007.htm

[6] N. F. Noy, D. L. McGuinness, and others, "Ontology development 101: A guide to creating your first ontology," 2001. [Online]. Available: http://protege.stanford.edu/publications/ontology_development/ontology101.pdf

[7] "Semantic web - w3c." [Online]. Available: http://www.w3.org/standards/semanticweb/

[8] "SNOMED CT." [Online]. Available: http://www.ihtsdo.org/snomed-ct/

[9] "The open biological and biomedical ontologies." [Online]. Available: http://www.obofoundry.org/

[10] D. Dzemydiene and E. Kazemikaitiene, "Ontology-based decision support system for crime investigation processes," in *Information Systems Development*, O. Vasilecas, W. Wojtkowski, J. Zupani, A. Caplinskas, W. Wojtkowski, and S. Wrycza, Eds. Springer US, 2005, pp. 427–438. [Online]. Available: http://dx.doi.org/10.1007/0-387-28809-0_37

[11] H. Park, S. Cho, and H.-C. Kwon, "Cyber forensics ontology for cyber criminal investigation," in *Forensics in Telecommunications, Information and Multimedia*. Springer, 2009, pp. 160–165. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-02312-5_18

[12] M. C. Surez-Figueroa, "D5. 4.1. NeOn methodology for building contextualized ontology networks," 2014. [Online]. Available: http://www.neon-project.org/deliverables/WP5/NeOn_2008_D5.4.1.pdf

[13] "OWL web ontology language overview." [Online]. Available: http://www.w3.org/TR/owl-features/

[14] "NEPOMUK representational language (NRL)." [Online]. Available: http://www.semanticdesktop.org/ontologies/2007/08/15/nrl/

[15] "RDF schema 1.1." [Online]. Available: http://www.w3.org/TR/2014/REC-rdf-schema-20140225/

[16] M. Vlkel, *Personal knowledge models with semantic technologies*. BoDBooks on Demand, 2010. [Online]. Available: http://digbib.ubka.uni-karlsruhe.de/volltexte/documents/1453712

[17] "VocabularyMarket - w3c wiki." [Online]. Available: http://www.w3.org/wiki/VocabularyMarket

[18] "OSCAF ontologies." [Online]. Available: http://www.semanticdesktop.org/ontologies/

[19] "Nepomuk contact ontology (NCO)." [Online]. Available: http://www.semanticdesktop.org/ontologies/2007/03/22/nco/

[20] "Nepomuk information element ontology (NIE)." [Online]. Available: http://www.semanticdesktop.org/ontologies/2007/01/19/nie/

[21] "GeoNames ontology - geo semantic web." [Online]. Available: http://www.geonames.org/ontology/documentation.html

[22] "Time ontology in OWL." [Online]. Available: http://www.w3.org/TR/owl-time/

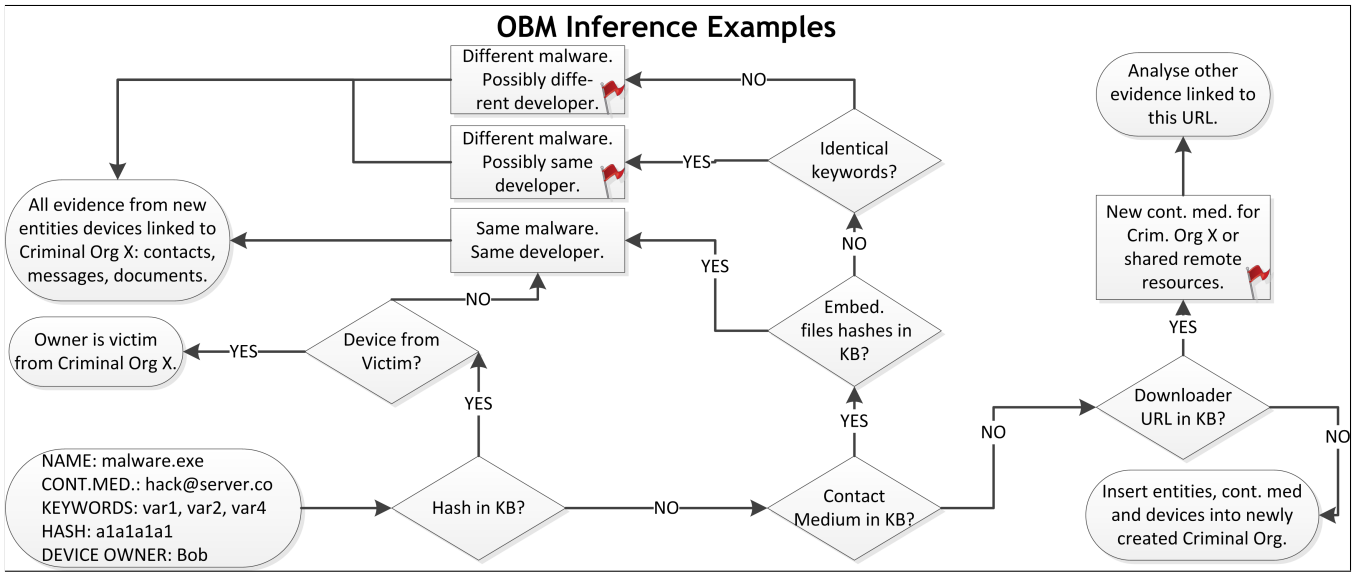


Fig. 4: Possible inference rules regarding malware files hashes.

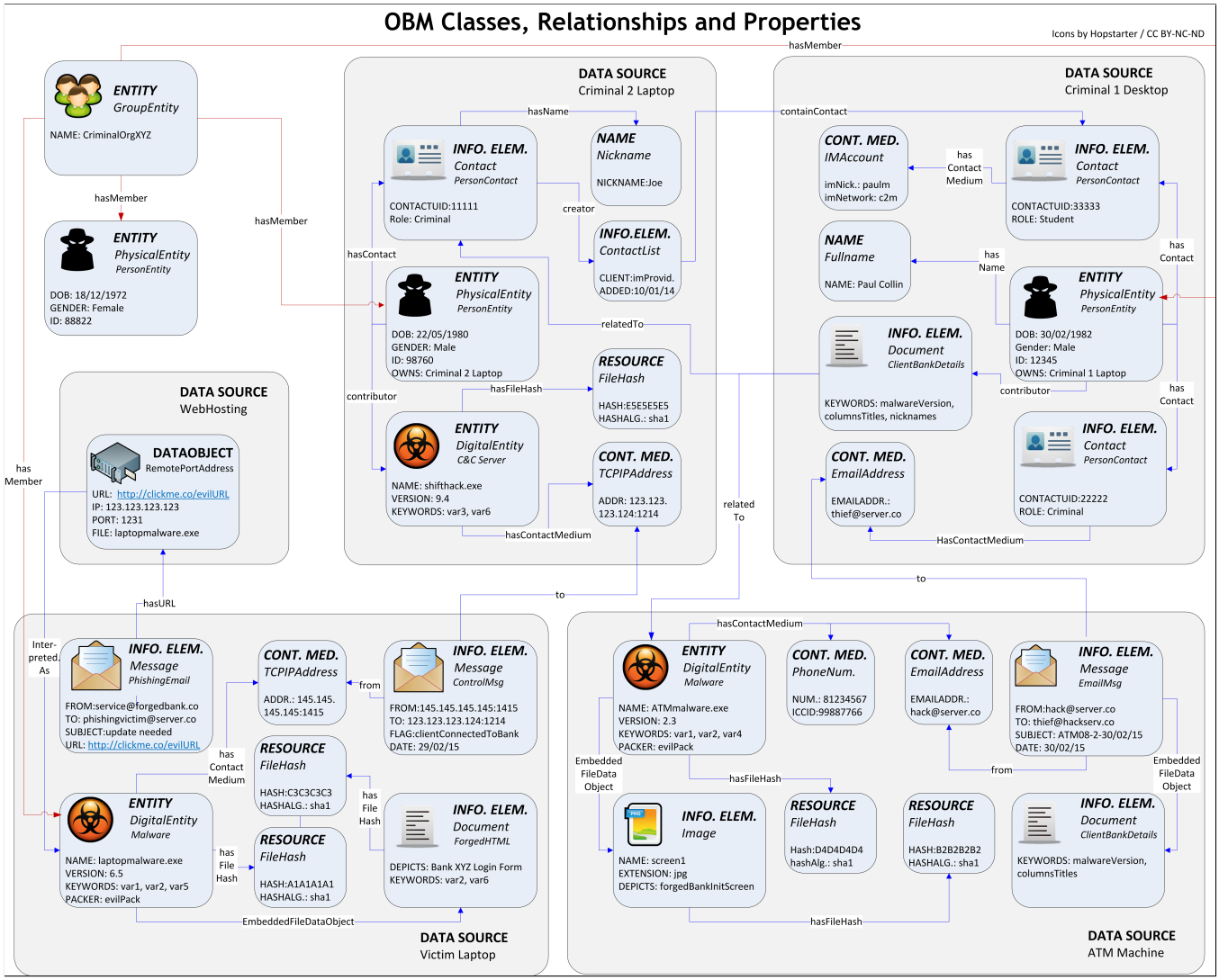


Fig. 5: Suggested concepts for the OBM domain.