A Description Logic Ontology for Email Phishing

Franklin Tchakounté*‡, Djeguedem Molengar*, Justin Moskolaï Ngossaha **

* Department of Mathematics and Computer Science, Faculty of Science, University of Ngaoundéré, Cameroon
** Department of Mathematics and Computer Science, Faculty of Science, University of Douala, Cameroon

‡ Corresponding Author; Tel: +237696465767, e-mail: tchafros@gmail.com

ORCID ID: 0000-0003-0723-2640, 000-0001-8411-2170, 0000-0002-9228-4543

Research Paper Received: 07.08.2019 Revised: 10.01.2020 Accepted: 17.02.2020

Abstract- Phishing detection is an area of identifying malicious activities designed by phishers to lure users providing sensitive information. Existing anti-phishing systems use blacklists based on specific parameters, characterize attacker's activities with artificial and computational approaches and educate users. The development and maintenance of these systems is hard and costly because of the polymorphic nature of phishing techniques. Phishing attacks are able to scam humans with insufficient knowledge, while countermeasures focus on specific characteristics to make decisions. Defining formal approaches for representing and reasoning knowledge in anti-phishing systems is therefore a concern. This work deals with this issue by proposing formalized description logic to build the knowledge base of phishing attacks. It additionally designs an ontology-oriented approach to add semantics on that knowledge. The ontology model has been proven consistent and satisfiable. Experimentations on case studies demonstrate the ability of the proposed model to represent knowledge attack scenarios. A comparison with state-of-the-art researches shows that the proposed formalism is more adequate to characterize phishing semantics. This work could successfully complement anti-phishing systems.

Keywords- Description logic; Ontology; Phishing.

1. Introduction

A knowledge-based system is a program able to reason to solve a certain problem, with the help of knowledge related to several fields such as medicine, knowledge engineering and cyber security [1]. Domain knowledge is represented by entities that have a syntactic description to which semantics are associated [2]. There is no universal method for designing knowledge-based systems, but research proposals has been developed around the logic of predicates, semantic networks, and frame languages [3]. They gave birth to a family of representation languages called Description Logics (DL), or terminological logics [4]. Ontologies represent conceptual models that transform knowledge specified by a language such as DL, in an exploitable form by an information system [5]. They are used to reason objects of the studied domain.

Phishing is a serious concern in the cyberspace[6][7]. It consists to identify malicious activities built by attackers to lure users, whom provide sensitive information such as bank account number, password, and so on [8]. It mainly relies on social engineering, to insert malicious URLs, counterfeit email, impersonate source and destination headers, and attached malicious files [9]. Phishing includes e-mail phishing which counterfeits an e-mail and URL phishing which designs fake URL to redirect users to the attacker. Literature mainly provides with two categories of anti-phishing attempts [10][11]. The first one includes detection approaches based on information-related characteristics ([12]–[20]) and prevention approaches based on user education

techniques ([21]–[24]). This category is limited since authors rely on static email features, misused in sophisticated techniques. Additionally, educative techniques are inconsistent due to ignorance of users and sophisticated social engineering techniques exploited. The second category includes approaches based on knowledge representation [25]–[32]. Literature offers just few works belonging specifically to this category. Authors who deal with phishing are restrictive to one scenario and none of them proposes any description logic representation. Unlike, this work proposes an ontology based description logic to characterize possible email phishing scenarios. It starts by providing a generic taxonomy of email phishing processes. It then designs related Tbox and Abox knowledge based on DL. This work builds an ontology related to this DL and successfully proves its consistence by the reasoner Racer. Several scenarios of phishing are matched to this ontology, to demonstrate its reliability.

This work is organized as follows: The first section reviews phishing countermeasures. The second section presents concepts about knowledge representation and phishing. The third section includes the model of DL, the construction of ontology and its reasoning. The fourth section experiments the ontology on some case studies of phishing email and makes a comparison with similar works. The next section concludes and gives research perspectives.

2. Related works

This section describes various orientations provided in literature to deal with phishing.

2.1 Detection-based approaches

This category includes various approaches. Authors rely on content and metadata to profile phishing traces. Similarity measures are designed to compare normal web pages and malicious pages [12], [33]. Machine learning approaches are applied to classify between benign and fake e-mails [13]–[20]. White and blacklisting are exploited to block malicious DNS entries [34]–[36].

2.2 Education-based solutions

Several authors propose solutions to enhance user awareness of different attacks and their technique [21]–[24].

2.3 Knowledge representation-based solutions

Some authors oriented their research towards formalizing objects, entities and their relationships in cybersecurity area.

2.3.1 Cybersecurity in general

Sikos [37] [Handling Uncertainty and Vagueness in Network Knowledge Representation for Cyberthreat Intelligence] proposes description logic representations of network knowledge originating from diverse sources to enable efficient automation via reasoning and to catch uncertainty and impreciseness. Ellison et al. [38] formalize description logics to represent and reason knowledge in digital forensics and digital security. Scarpato et al. [39] couple description logics to Web Ontology Language (OWL) and SPARQL queries able to represent information needed to generate the Reachability Matrix withinin the Open Systems Interconnection(OSI) protocols. They deduce ontology for cybersecurity.

2.3.2 Anti-phishing proposals

Literature provides very few anti-phishing techniques which include ontology and DL representations. Tseng et al. [25] propose an ontology based on the framework language to model and represent a phishing attack scenario. Bazarganigilani [26] proposes an ontology model for semantics on a phished email classifier based on the Naive Bayes algorithm. Qaseem and Govardhan [27] propose a system to catch phishing in instant messaging. For that, they designed an ontology to represent the context or intention through objects related to instant messages exchanged between chatters. Kerremans et al. [28] propose a knowledge representation architecture to discriminate various types of scams including phishing. Zhang et al. [32]

propose a phishing domain ontology representing linguistic characteristics of page contents. Authors rely on this information to look if a website is similar to the ontology of phishing website. Kiran et al. [30] aim to check webpage's legitimacy by understanding its content. For that, they build RDF of nineteen elements of webpages to rely on for identifying the nature of a webpage. Park and Rayz [40] reinforce the classifiers by adding ontological semantics to terms exploited in emails and websites. Falk [31] focuses on just the language used in the bodies of email messages to model ontological objects and relations within email contents. He demonstrates that machine learning algorithms are improved.

2.4 *Comparison*

Table 1 compares the aforementioned phishing category proposals.

Table 1. Phishing proposals comparison

Proposal	Advantages	Disadvantages
Detection-based	This category prevents the appearance of a phishing site	Hackers can easily deceive by providing them with a visually similar site. The rate of false positives and false negatives: these solutions are based on the characteristics related to the object: e-mail, files, URL. They are limited due to the polymorphic nature of phishing.
Educative-based solutions	This category improves awareness and increases knowledge about attacks	Hackers use social engineering methods to lure users. By nature, users trust each other. Attacks are evolving whereas users are not in their training.
Knowledge representation-based solutions	This category defines properly the semantic relationships among domain concepts. It increases the likelihood of detecting new forms of phishing e-mails [41].	There are many objects to represent. It could be extremely time consuming while respecting the constructors of the representation formalism. Works dealing with phishing are rare and existing ones lack to propose generic ontology based DLs. They are too restrictive to one scenario.

2.5 Summary

This research provides a generic ontological model based on DLs to model phishing knowledge since it could be complementary to detection-based and educative-based systems.

3. Background

This section presents concepts related to knowledge representation and phishing.

3.1 Knowledge representation

Knowledge relies on cognitive schema specific to each individual. One of the challenges in Artificial Intelligence (AI) consists to model a domain and to implement this cognition in a form exploitable by both humans and by machine.

3.1.1 Semantic networks

According to Quillian's works [42], semantic networks are originally designed to illustrate the memory processing in the context of linguistics. They become later a language of representation [8]. Despite the contribution of semantic networks in the field of knowledge

modelling, they have significant limitations. In fact, they focus on information structure and not on its semantics. This situation could lead to confusion between relations or classes [43]. The previous observation leads to the development of new formalisms called frames.

3.1.2 Schemas (frames)

According to Minsky [44], the principle of frame relies on the assumption that the representation of the World is in the form of arrangement of elementary information units called schema. Unlike semantic networks, for which semantic memory is associative, information can be divided into subsystems that potentially have inter-links. The idea is to collect all the necessary information about a situation and to put them in a place, called frame. Some pioneers such as Hayes [43] criticized the absence of formal semantics in this formalism.

3.1.3 Conceptual graphs

Inspired by the existential graphs of Pierce and semantic networks, Sowa [44] provides a new mode of representation based on predicate logic of first order. They are used in database structure. The conceptual graph can be connected, finite and bipartite graph. Like frames, they lack expression of semantics in the definition of relations and classes [45].

3.1.4 Description logic

Previous representations have been formalized using logic to give semantics. Without such a precise formalization, they are vague and ambiguous, and thus problematic for computational purposes. Description logics extend frame-based systems by expressing definitions of classes and relations [46]. Several description logic languages exist and differ in language expressiveness. DL languages provide formal semantics and can therefore represent the knowledge of an application domain in a structural and formal way. DL is used in this research because of several reasons.

- Description logics have become a major knowledge representation paradigm, in particular for use within the semantic Web. It can be applied in cyber security [21]–[23], [27], [28], [47]–[49]
- DL is decidable, i.e. given a concept, it is possible to determine if this definition is consistent with others. Also given an instance definition, it can be decided which is the concept definition that most fits it.
- DL has sound and complete reasoning mechanisms which guarantee the results accuracy and reliability.
- Wide range of logics has being developed till now, from very simple (no disjunction, no full negation) to very expressive, so logic satisfying research needs could be selected in a minimum computational complexity.
- Modern DL reasoning engines are quite efficient when providing results.

DLs rely on three notions such as concept, role and individual [50]. Concepts correspond to classes of individuals, roles are relationships between these individuals, and individuals correspond to individual concepts. In a descriptive logic knowledge base, there are two components: TBox and ABox. The first contains all the axioms defining the concepts of the domain. ABox contains assertions about individuals, specifying their classes and their attributes.

3.2 DL families

DLs have a common base language called Attributive Language (AL). It is enriched with the following syntactic elements [51]:

- A: Atomic concept
- $\top$: Universal concept Top
- $\bot$: Empty Bottom concept
- $\neg A$: Negation of atomic concept
- $C \sqcap D$: Conjunction of concepts
- $\forall$ r.C : Universal quantifier
- $\exists$ r : Existential quantifier non-typed

Some expressive languages can be derived by adding other constructors to the AL language.

- ALU = AL ∪ {C ⊔ D} : disjunction of concepts;
- AL∈ = AL ∪ { ∃ rC } : existential typed quantifier ;
- Attributive Language with Complement (ALC): this is the most important logic since it is the basis of all logics of expressive description;
- ALC = AL ∪ {¬C}: Here, C is a primitive or defined concept;
- ALN = ALU {≤nr, ≥nr}: Number restrictions, denoted by the letter N and denoted by ≤ nr (restriction to less than n) and ≥ nr (restriction to more than n) or n represents a positive integer.

AL can be extended using constructors of concepts and constructors of roles.

3.3 Satisfiability of a concept

Definitions: Satisfiability of a concept, equivalence, incompatibility

• A concept C is *satisfiable* or coherent if and only if there is an interpretation I such as $C^I \neq \emptyset$; It is unsatisfiable or inconsistent otherwise.
• Two concepts C and D are said to be *equivalent*, which is noted C ≡ D, if and only if
$C^I = D^I$ for any interpretation I.
• Two concepts C and D are *incompatible* or *disjoint* if and only if $C^I \cap D^I = \emptyset$; for any interpretation.

3.4 Knowledge base

In description logic approaches, the representation of knowledge includes two levels: TBox, which allows to reason only on concepts, and ABox, which introduces reasoning on individuals. ABox includes a set of assertions about individuals, such as assertions of memberships and role assertions. As shown in Figure 1, the knowledge base (KB) relies on a language and can be implicitly enriched using inference models. It can be exploited by information systems to improve decision making.

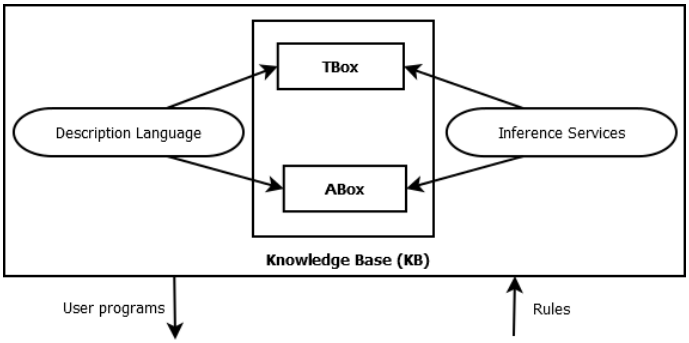


Fig. 1. General structure of an DL system [51]

3.5 Reasoning

Humans process knowledge by reasoning so that they reach conclusions [52]. Analogously, a computer processes the knowledge stored in KB by drawing conclusions from it, i.e., by deriving new statements that follow from the given ones.

3.5.1 Subsumption

The subsumption relation used in the knowledge base (Tbox, Abox) is the relation (is-a) which is the relation of subsumption with inheritance [53]. It is only used between classes of the same ontology and allows defining simple hierarchies.

3.5.2 Reasoners

Reasoners check for logical contradictions and for consistency of ontology model [54]. A reasoner can invalidate ontologies in different ways:

- An ontology can be detected as inconsistent meaning that there is no possible interpretation of the ontology;
- An ontology is insatisfiable when there is a possible interpretation of that ontology.

3.5.3 Inference Engine

The inference engine [55] is the set of reasoners for inferring on the basis of knowledge. It is a program that makes the logical deductions of an expert system from a knowledge base and a database of rules. In fact, most of these engines are designed to reason on the description logic, but accept Web Ontology Language (OWL) files as inputs. Once the

ontology is loaded, these engines make inferences about TBox and ABox.

3.6 Ontologies

The notion of ontology was first addressed by John McCarthy in the field of artificial intelligence (AI) [56]. An ontology is simply the set of concepts, relations, attributes and hierarchies existing in a domain [5].

3.6.1 Benefits

Ontologies are helpful in several points.

- It provides a common understanding of the information structure between people and software manufacturers;
- It renders interoperability between systems;
- It facilitates exchange of knowledge between systems;
- It facilitates reuse of knowledge on a domain by creating and maintaining reusable knowledge bases.

3.6.2 Representations

An ontology can be represented as a graph of concepts and relations (graph of ontology), as a model to formally describe resource on the Web (Resource Description Framework) and as description logic notations This work couples the graph representation to DL notation since RDF is more oriented programming.

3.6.3 Type of ontologies

There are four main types of ontologies. Top-level ontologies which describe abstract and general concepts that exchanged across various domains and applications. Domain ontologies capture the knowledge within a specific domain of discourse such as phishing. Task ontologies capture the knowledge within a specific task, such as phishing analysis. Application ontologies which are within the scope of this work combine both domain and task ontologies.

3.7 Phishing

Phishing is a cyber-criminal technique exploiting social engineering to lure target users providing sensitive information such as bank accounts. This document deals with spear phishing characterized by the fact that phishers designed a fake e-mail targeting a specific group of individuals. According to [11], phishers mainly follow four steps.

- Designing and dissemination of fake e-mails during which attackers design fake e-mails and flood them through messaging means to the targeted users.
- Visiting malicious websites during which the victim is redirected to the phisher website.
- Releasing of sensitive information during which the user is persuaded to disclose confidential information.
- Gathering of sensitive information during which user’s confidential information are sent to attackers.

In Figure 2, a more specific and complete phishing process is proposed. It includes collecting information on the target, representation of elements of incitement and phisher exploits.

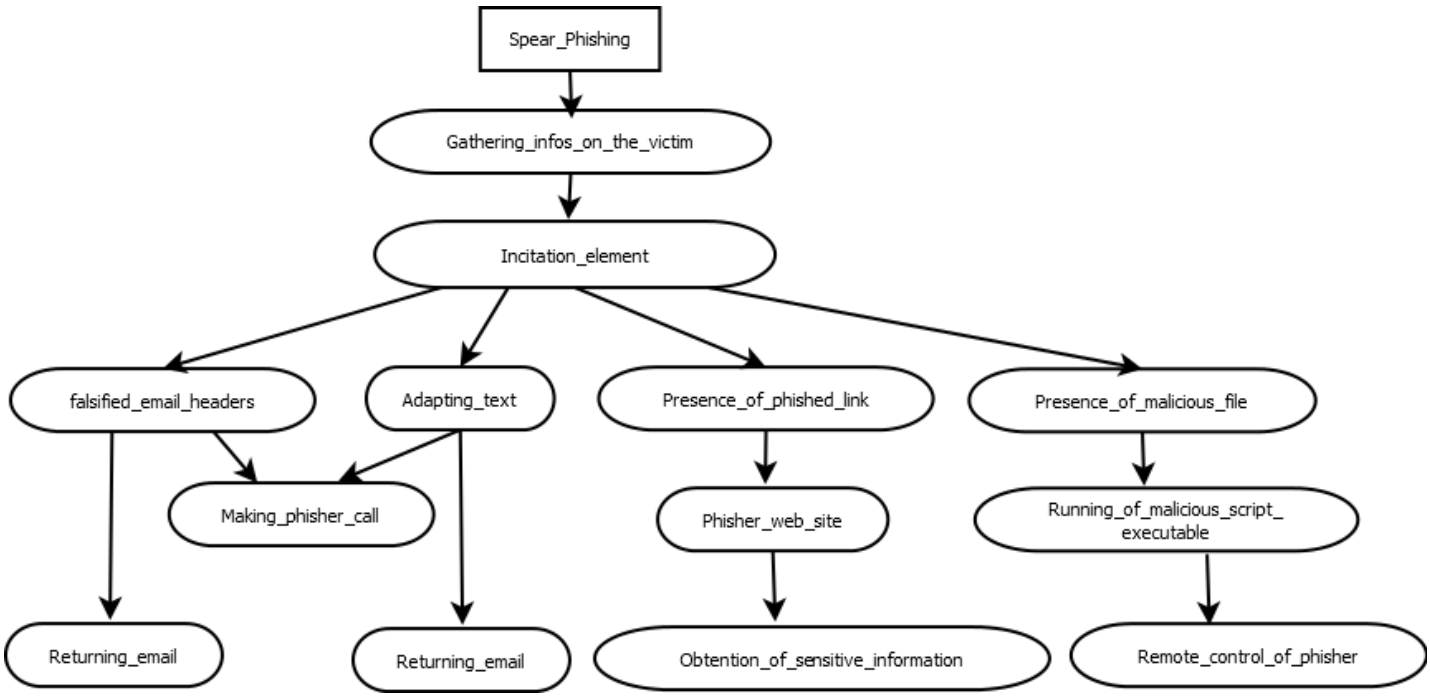


Fig. 2. Scenario of spear phishing

The phisher begins with a sharp investigation of the target. It can be done either through social media or by collecting personal traces left on the Web. Once the collection is sufficient, the phisher prepares the phishing attack by minutely designing mail contents and falsifying the mail header. This falsification aims at usurping the victim identity. For example, the source address used by the hacker can be me@yah00.com instead of me@yahoo.com which corresponds to the real address of the user. The phisher can also fudge a text by masquerading as a well-known body by adding emergency words to incite the victim to act. The hook may be to make a phone call to the recipient or to click on a bad link redirecting the user to a malicious site. Once the site is opened, counterfeit forms collect sensitive information and malicious scripts are launched in background to remotely take control. Another method is to attach malicious files to emails. The phisher at this point uses social engineering tricks through words or pictures to get the victim to click on the file to launch it. At this point, thanks to the malicious scripts attached to the file and the vulnerabilities of the client software needed to read it, the phisher can have remote control of the host computer.

4. Description logic model

The knowledge base includes Tbox and Abox. Their specification relies strictly on Figure 2 which describes phishing scenarios. The following section specifies KB of the phishing domain.

4.1 TBox

Figure 3 shows the TBox configuration space for phishing attacks. It starts by formalizing the *Person* axiom. This axiom can be either the victim or the phisher (1). The *Phisher* axiom is defined as a person who targets a victim to deceive (2), and a *Victim* as a person who has been deceived (3). Spear phishing is subsumed by the intelligence about gathering information about the victim (target) and the incentive element (4). The description of the latter logics relies to the scenarios in Figure 2.

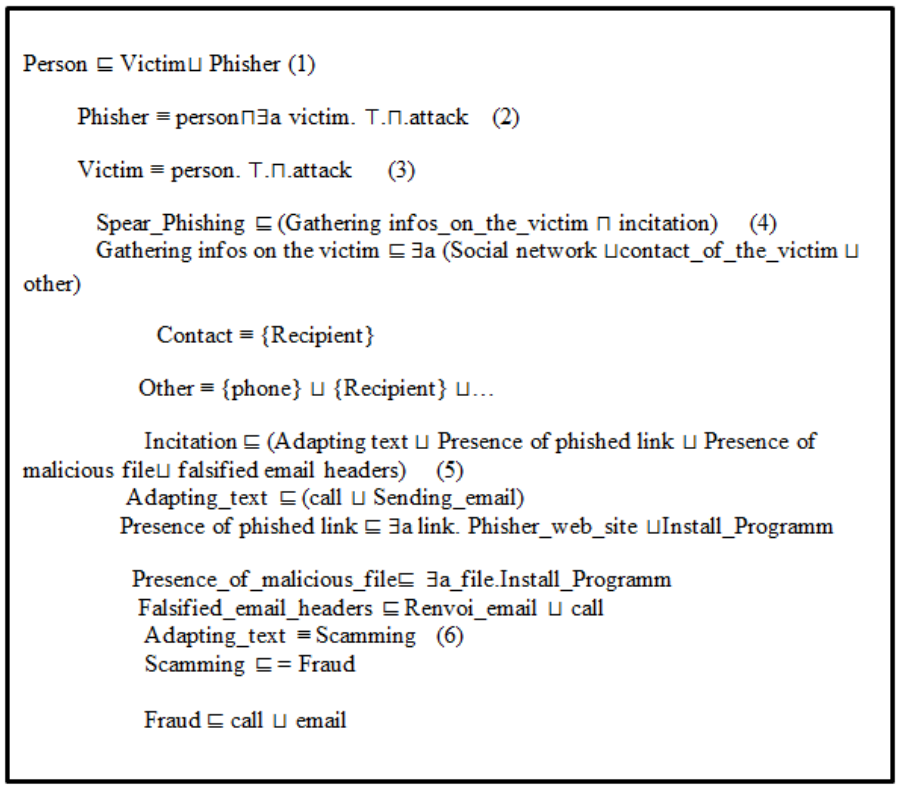


Fig. 3. TBox

4.2 ABox

ABox contains the key elements of TBox.
A = {(i: incitation, p: phisher, v: victim, ta: adapted_text, has_an adapted_text (v; p), cd: Information_collection, f: form, a_form_ (cd; f), io: computer_infection_, lm: a_lien_lien_lien_ (io; l m), pj: a_an_table (io; pj)}
Here *i; p; v; your; cd; f; l m; io; pj; pr; cv*; *am* are individuals of the following concepts: *incentive, phisher, victim, adapted_text, Collection_information, form, infect_computer, attach_party, take_information, contact_of_the_victim, other_means* respectively. The knowledge base should be tested for consistency and, therefore, the user's choice will be validated.

4.3 Construction of the ontology

4.3.1 Construction

Protégé 4.3 is used to build the ontology through Ontology Web Language (OWL) [59]. We adopt OWL DL because, firstly, the OWL DL makes it possible to express multiple cardinalities and, on the other hand, the other languages are unsatisfactory or more complex. Resource Description Framework (RDFS) is limited since it does not allow the expression of cardinality constraints.
Protégé 4.3 requires the installation of the Owlapi 4.2.0 Application Programming Interface (API) required for the manipulation of the ontology. The ontology is obtained after the following phases:

- The launch of Protégé;
- The creation of classes and subclasses;
- The creation of properties.

The ontology contains three (03) main classes as shown in Figure 4.

- The class "spear_phishing" includes all the other classes because it is from it that the class intelligence on the target arises. The class element of incitement, “element_incitation”, contains the subclasses which are among others: Information on the falsified header, adapted text,

presence of a phished link, presence of a malicious file containing also subclasses.

- The subclass shadow link contains the following elements:
    - Running a malicious program;
        - Phisher's site;
        - Remote control of the host machine.
- The malicious subclass contains the following:
    - Running a malicious program;
    - Collection of personal or confidential information;
    - Remote control of the host machine.

Figure 5 illustrates the ontology for phishing attacks based on DL. It has been found that there are two (02) arrows on both sides that point to classes and subclasses, the purple arrow is a "contained" relationship and the second is the class property. It appears or it points to the corresponding class, meaning that the incentive element can contain either a suitable text or a malicious file, or information on the header. When it is a falsified or a phished link, it has the same meaning (i.e. the relation contains) as the first. But the latter allows seeing the property between classes and subclasses.

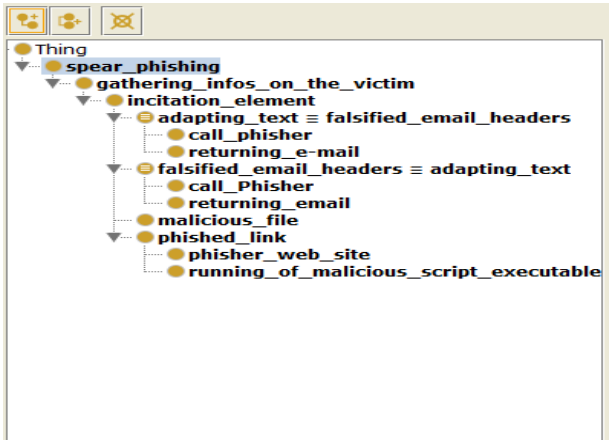


Fig. 4. Interface for class naming

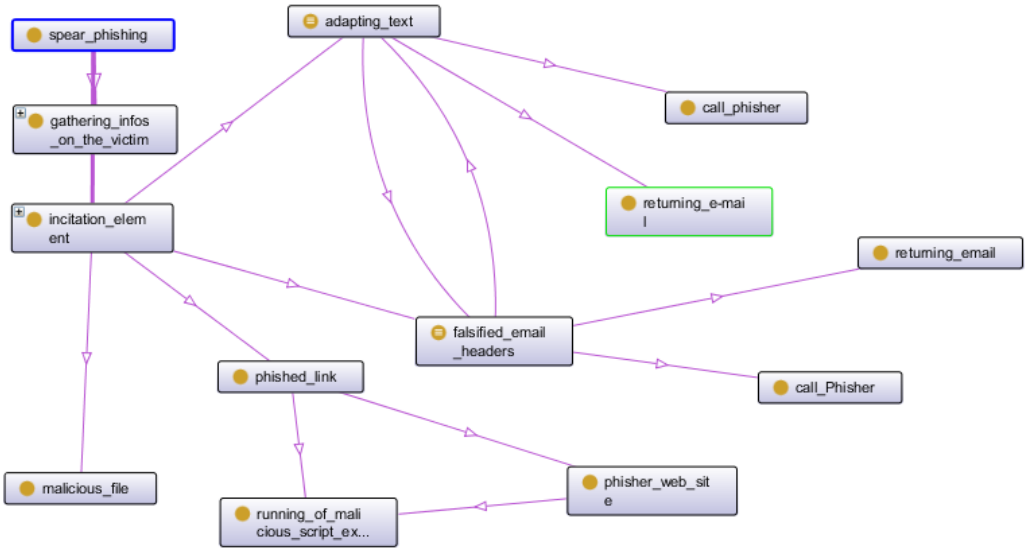


Fig. 5. Ontology for phishing attacks.

4.3.2 Inference

The generation of inferences is done via a query language called Simple Protocol and RDF Query Language (SPARQL). This language acts on the data stored in RDF. The loaded ontology requires OWLAPI which is an open source Java API. It is used to manipulate ontologies and generates .owl files representing the ontology. OWLAPI takes care of analyzing the file, extracting the axioms and creating the OWL Ontology class.

4.4 Reasoning

Manual and automatic reasoning are applied to check the consistence of the ontology.

4.4.1 Manual reasoning

Manual reasoning is performed by using the tableau reasoning [60]. It is the best method for making inferences with the description logic [61]. The goal is to reason phishing attack DLs with the reasoning by table to demonstrate the satisfiability of the proposition representing phishing attacks as defined in Equation (1).

$$\text{Phishing} = \forall\ (\leq 1 \text{Dis. Link.Mal} \sqcap \text{Site.Visit} \sqcap \text{Info.Sensitive.Releve} \sqcap \text{Transfer.Info. Disclose}) \quad (1)$$

Equation (1) is used as starting point a predicate independent of any terminology meaning with TBox excluded. It is done by replacing all the terms of the formula by their definition in the terminology. Indeed, if a term in the formula has no definition in terminology, it remains unchanged. This process is repeated until the formula obtained contains no term that has a definition in the terminology.

The black square in Figure 6 refers to a contradiction. The proof is represented by a tree, where each branch represents alternatives. After repeated substitution, alternatives provide contradiction as shown in Figure 6.

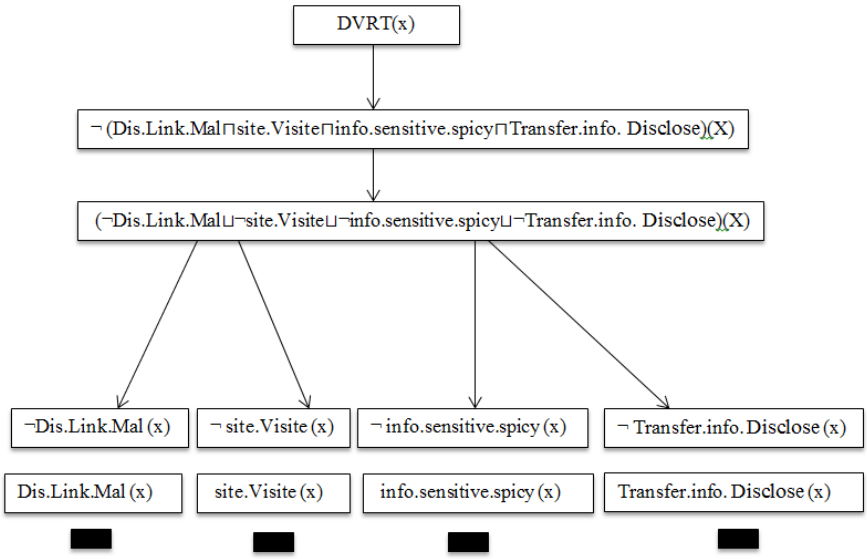


Fig. 6. Reasoning Tree

4.4.2 Automatic reasoning

Verifying the consistency of this ontology consists of checking its satisfiability. It is realized using the code shown in Figure 7.

```
public class Test1 {
    public void loadOntology() throws OWLException {
        OWLOntologyManager om = OWLManager.createOWLOntologyManager();
        File file = new File("C:\\Users\\Molengar Fils\\Documents\\owl\\ontol
        OWLOntology MonOntologie = om.loadOntologyFromOntologyDocument(file);
        System.out.println("Mon ontologie est:\n " + MonOntologie);
        OWLReasoner reasoner;
        OWLReasonerFactory reasonerFactory = new StructuralReasonerFactory();
        ConsoleProgressMonitor progressMonitor = new ConsoleProgressMonitor()
        OWLReasonerConfiguration config = new SimpleConfiguration(progressMon
        reasoner = reasonerFactory.createReasoner(MonOntologie, config);
        reasoner.precomputeInferences();
        boolean consistent = reasoner.isConsistent();
```

Fig. 7. Verification of ontology consistency

```
    ... finished
Consistent: true

------------------------------------------------------------------------
BUILD SUCCESS
------------------------------------------------------------------------
Total time: 4.431s
Finished at: Wed Dec 20 04:42:42 CET 2017
Final Memory: 11M/106M
------------------------------------------------------------------------
```

Fig. 8. Consistency check outputs

Figure 8 shows that the proposed ontology is consistent.

5. Experimentation and interpretation

The spear phishing ontology is tested and proved consistent. Here, we considered five experimental scenarios: email with an incentive element, email with information on the falsified header, e-mail with a fake content and phished link, email with malicious attachment, and email with grammatical mistakes with malicious attachment.

5.1 Email with an incentive element

This experiment checks whether the constructed ontology respects the practical case of spear phishing illustrated in Figure 9. It is a falsified Paypal message aiming to incite the victim to click on the confirmation button to be redirected into the phisher's site. The proposed ontology is able to characterize this scenario since it includes a concept about incitation. In this case, it concerns the simulated logo and the sentence "Your account PayPal is limited. You Have To Solve The Problem in 24 Hours" urging the user.

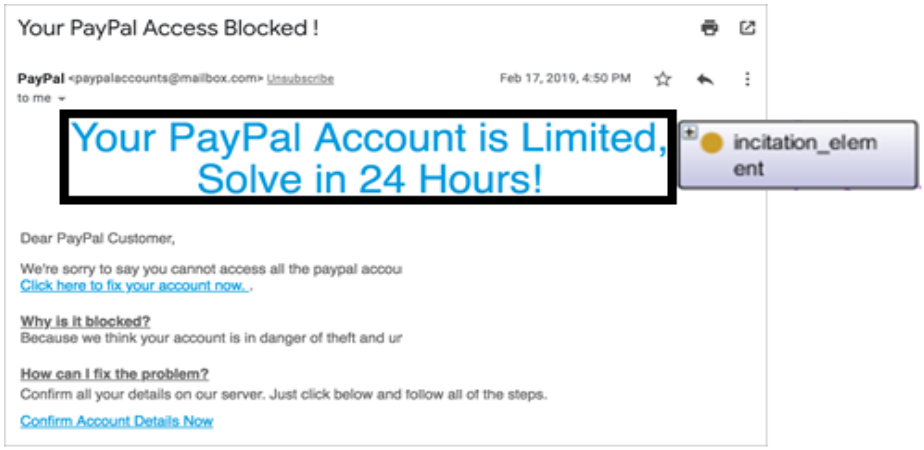


Fig. 9. Email with incentive element.

5.2 Email with information on the falsified header

Figure 10 shows a phished email including a falsified sender address. In fact, the phisher builds this address using as prefix the name of real contact and suffix(after the "@") a modified domain name. This scenario belongs to the ontology as falsified email headers. But fake headers could be seen as incitation element, also well described by the ontology.

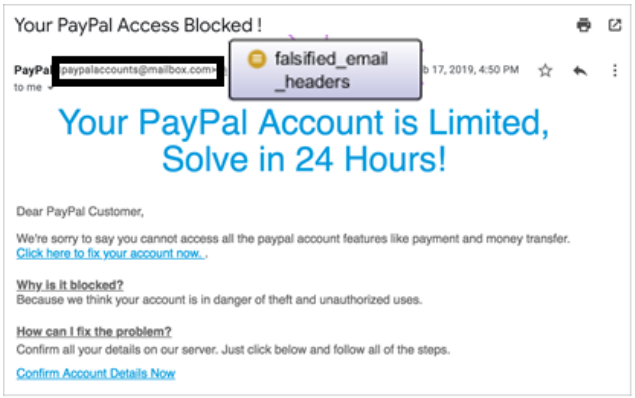


Fig. 10. Email with information on the falsified header

5.3 Email with adapted text and phished link

Figure 11 shows a message that informs that a PayPal account has been limited, and that there is a short time to solve the problem causing this limitation. And this email contains a phished link, which redirects to a fake site (site of the phisher). The hacker's site presents a form to provide sensitive information. All these three elements (adapted text, phished link and the pirate's site) exist in the created ontology.

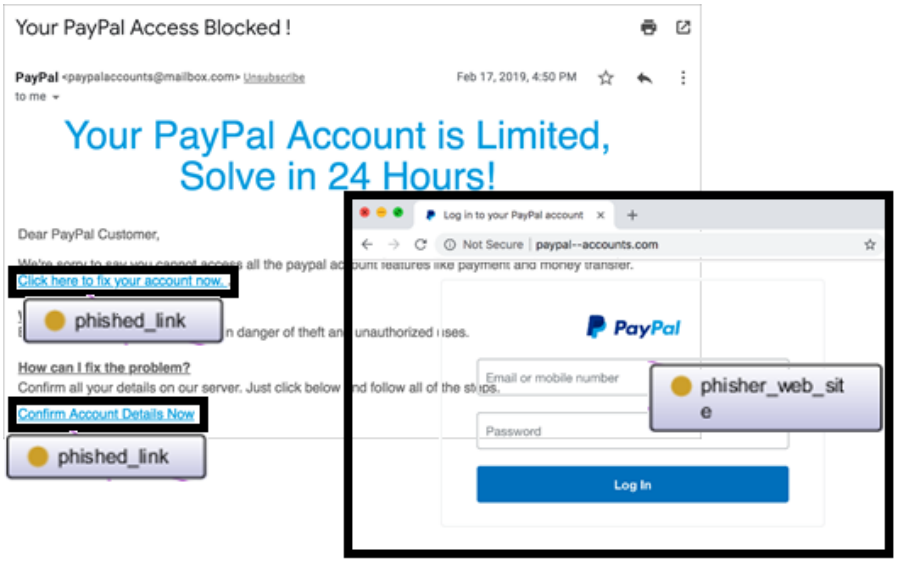


Fig. 11. Email with adapted text and phished link

5.4 Email with malicious attachment

Figure 12 illustrates the receipt of an email containing a fake attachment named invitation.htm rather than a .pdf file. Once clicked on the piece attached, it redirects to a fake Gmail website which is the phisher's site. The proposed ontology proposes concepts to characterize malicious attachment as well as phished website knowledge.

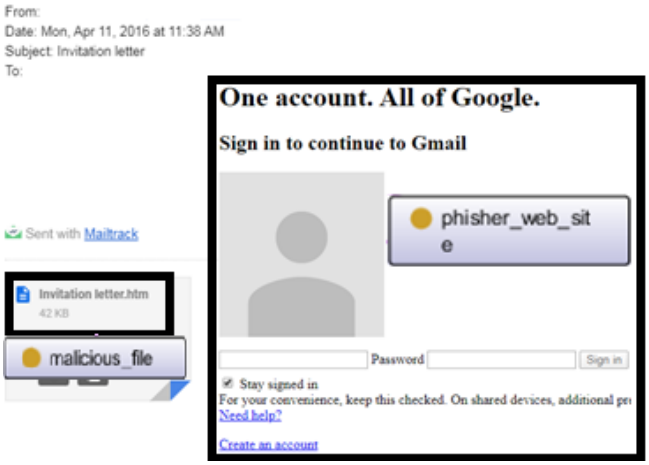
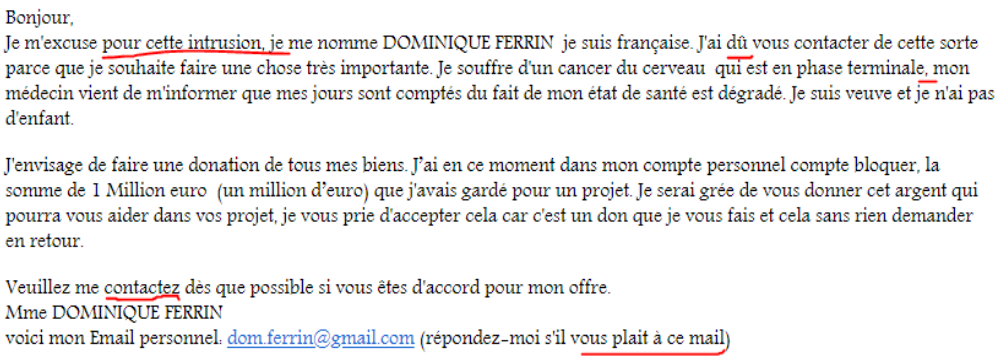


Fig. 12. Email with attachment

5.5 Email with grammatical mistakes and a malicious attachment

The received email (shown in Fig.13) is a fake one, identified by the grammatical, vocabulary and punctuation mistakes. Unfortunately, the ontology is unable to deduce knowledge from this scenario because the language used for the ontology does not include artificial intelligence dealing with sentence mistakes. The language used in this research is descriptive.



Bonjour,
Je m'excuse pour cette intrusion, je me nomme DOMINIQUE FERRIN je suis française. J'ai dû vous contacter de cette sorte parce que je souhaite faire une chose très importante. Je souffre d'un cancer du cerveau qui est en phase terminale, mon médecin vient de m'informer que mes jours sont comptés du fait de mon état de santé est dégradé. Je suis veuve et je n'ai pas d'enfant.

J'envisage de faire une donation de tous mes biens. J'ai en ce moment dans mon compte personnel compte bloquer, la somme de 1 Million euro (un million d'euro) que j'avais gardé pour un projet. Je serai grée de vous donner cet argent qui pourra vous aider dans vos projet, je vous prie d'accepter cela car c'est un don que je vous fais et cela sans rien demander en retour.

Veuillez me contactez dès que possible si vous êtes d'accord pour mon offre.
Mme DOMINIQUE FERRIN
voici mon Email personnel: dom.ferrin@gmail.com (répondez-moi s'il vous plait à ce mail)

Fig. 13. Email with grammatical errors and an attachment

5.6 Comparison with similar works

Table 2 presents some criteria to evaluate similar researches against the proposed scheme in this document. Five criteria are exploited. The first criterion refers to the style used to represent the ontology. The second criterion indicates whether the ontology is top-level, domain, task or application as described in section 3.6.3. The third criterion gives the reason why the ontology has been used in the proposal. The fourth criterion indicates scenarios among those experimented in this work, which are identifiable through the proposals. The fifth column concerns some observations. It appears that most research exploits ontology during text analysis to semantically represent concepts related to terms used by phishers. Authors identify phishing pages or phishing emails by analysing linguistic features and to give semantic model to describe scenarios. Moreover, authors propose specific phishing domain ontology. It means that they only consider one aspect of possible techniques, i.e., analysing terms used in email content or page content. However, there are extended possible scenarios exploited by phisher to infiltrate as designed in this work (see Figure 2). Proposals represent knowledge with various schemes. The proposal exploiting description logics [32] aims to define semantic relationship of the word elements in the sentences. Unlike, our proposal models knowledge of up-to-date techniques of phishing based on description logic. We then associate a developed ontology to design knowledge usable by a computer system. An advantage of similar works is that they can directed be coupled to existing classifiers to improve detection of misclassifications. Most of them are domain-oriented although they capture only partial aspects of phishing. The proposed approach in this document is on contrary application-based meaning that it describes logics to represent the whole phishing domain but additionally build the ontology to make the knowledge exploitable in a system. Figure 13 reveals that 75% of similar works are able to identify the scenario 3 since they represent semantics related to terms in phishing terms. Unlike our proposal which is not able to detect scenario 5, works [27] and [28] do since they conceptualized lexical and grammatical terms in texts. One can rather develop ontology to represent frequent grammar and vocabulary mistakes in phishing messages.

Table 2. Summary of comparisons among similar works

	Representation of ontology	Type of ontology	Use of ontology	Scenario n° identified in the proposal	Observations
Qaseem and Govardhan [27]	RDF	Top-level	identify the semantic domain and context for the instant message keywords	5	specific
Tseng et al. [25]	Frame	Application	model a phishing attack scenario and related knowledge	3	coarse-grained
Kiran et al. [30]	RDF	Top-level	Model web elements of suspicious page	3	specific
Kerremans et al. [28]	Based on GOMA - RDF/OWL	Domain	Model concepts and relationships explicitly verbalized and related to lexicons	5	specific to email fraud
Falk [31]	Based on OntoSem – Frame and graph	Domain	Describe the language seen in the sample data	3	ontology support misclassifications
Park and Rayz [40]	Based on Ontological Semantic Technology	Domain	Represent syntax and semantics in terms of emails	3	ontology support classifications
Bazarganigilani [26]	undefined	Domain	Model different meanings and synonyms of terms	3	ontology support phishing page classifications
Zhang et al. [32]	DL	Domain	Define the semantic relationship of the word elements in the sentences appeared in the known phishing	3	ontology support phishing page classifications

			pages		
Our proposal	DL	Application	Represent knowledge of possible phishing scenario	1-4	generic

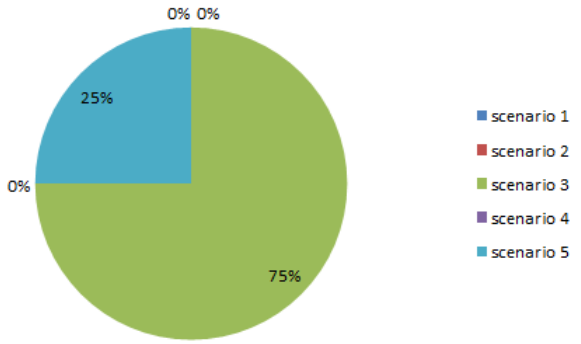


Fig. 13. Proportions of scenario

5.7 Limitations

The created ontology proposes a coarse-grained conceptualization. Indeed, basic knowledge describes general concepts without affecting details or values. This can be justified by the fact that the ALC language on which the ontology is based is expressive. The consequence is that detection based on grammatical errors has been found unsuccessful.

6. Conclusion and perspectives

In this work, we exploited description logics as the main support to design ontology to represent phishing knowledge. The knowledge base includes Tbox and Abox representing key elements to describe generic email phishing process. An ontology scheme is proposed based on DL representations. Ontology's satisfiability and consistency have been proven. It demonstrates that this ontology is exploitable in information systems and can be extended to new axioms.
As future work, a system exploiting the ontology for phishing education will be designed and implemented. The ALC language will be extended to more expressive DLs, by developing subsumption algorithms while optimizing complexity.

References

[1] A. Patel and S. Jain. "Formalisms of Representing Knowledge". *Procedia Comput. Sci.*, vol. 125, pages 542–549, 2018, doi: 10.1016/J.PROCS.2017.12.070.

[2] V. Nazaruks and J. Osis. "A Survey on Domain Knowledge Representation with Frames". *Proceedings of International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE)*, pages 346–354, 2017.

[3] B. Nebel. "Logics for Knowledge Representation". *Int. Encycl. Soc. Behav. Sci.*, pages 319–321, 2015, doi: 10.1016/B978-0-08-097086-8.43053-9.

[4] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider. "The Description Logic Handbook: Theory, Implementation and Applications, 2nd ed". Cambridge University Press, 2010.

[5] M. N. Asim, M. Wasim, M. U. G. Khan, W. Mahmood, and H. M. Abbasi. "A Survey of Ontology Learning Techniques and Applications". *Database*, vol. 2018, 2018, doi: 10.1093/database/bay101.

[6] D. Goel and A. K. Jain. "Mobile phishing attacks and defence mechanisms: State of art and open research challenges". *Comput. Secur.*, vol. 73, pages 519–544, 2018, doi: 10.1016/j.cose.2017.12.006.

[7] APWG. "Phishing Activity Trends Report 4th Quarter 2018". *Report*, 2019.

[8] M. Nicho, H. Fakhry, and U. Egbue. "When Spear Phishers Craft Contextually Convincing Emails". *Proceedings of International Conferences on WWW/Internet and Applied Computing*, 2018.

[9] F. Salahdine, N. Kaabouch, F. Salahdine, and N. Kaabouch. "Social Engineering Attacks: A Survey". *Futur. Internet*, 11(4), p. 89, 2019, doi: 10.3390/fi11040089.

[10] K. L. Chiew, K. S. C. Yong, and C. L. Tan. "A Survey of Phishing Attacks: Their Types, Vectors and Technical Approaches". *Expert Syst. Appl.*, vol. 106, pages 1–20, 2018, doi: 10.1016/J.ESWA.2018.03.050.

[11] A. Aleroud and L. Zhou. "Phishing Environments, Techniques, and Countermeasures: A Survey". *Comput. Secur.*, vol. 68, pages 160–196, 2017, doi: 10.1016/J.COSE.2017.04.006.

[12] R. S. Rao and A. R. Pais. "Two level filtering mechanism to detect phishing sites using lightweight visual similarity approach". *J. Ambient Intell. Humaniz. Comput.*, 2019, doi: 10.1007/s12652-019-01637-z.

[13] K. L. Chiew, C. L. Tan, K. S. Wong, K. S. C. Yong, and W. K. Tiong. "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system". *Inf. Sci. (Ny).*, vol. 484, pages 153–166, 2019, doi: 10.1016/j.ins.2019.01.064.

[14] R. S. Rao and A. R. Pais. "Jail-Phish: An improved search engine based phishing detection system". *Comput. Secur.*, vol. 83, pages 246–267, 2019, doi: 10.1016/j.cose.2019.02.011.

[15] M. Volkamer, K. Renaud, B. Reinheimer, and A. Kunz. "User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn". *Comput. Secur.*, vol. 71, pages 100–113, 2017, doi: 10.1016/j.cose.2017.02.004.

[16] S. W. Liew, N. F. M. Sani, M. T. Abdullah, R. Yaakob, and M. Y. Sharum. "An effective security alert mechanism for real-time phishing tweet detection on Twitter". *Comput. Secur.*, vol. 83, pages 201–207, 2019, doi: 10.1016/j.cose.2019.02.004.

[17] D. Delgado-Gómez, J. C. Laria, and D. Ruiz-Hernández. "Computerized adaptive test and decision trees: A unifying approach". *Expert Syst. Appl.*, vol. 117, pages 358–366, 2019, doi: 10.1016/j.eswa.2018.09.052.

[18] T. Nagunwa, S. Naqvi, S. Fouad, and H. Shah. "A Framework of New Hybrid Features for Intelligent Detection of Zero Hour Phishing Websites". *Advances in Intelligent Systems and Computing*, 2020, vol. 951, pages 36–46, doi: 10.1007/978-3-030-20005-3_4.

[19] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri. "Machine learning based phishing detection from URLs". *Expert Syst. Appl.*, vol. 117, pages 345–357, 2019, doi: 10.1016/j.eswa.2018.09.029.

[20] V. Patil, P. Thakkar, C. Shah, T. Bhat, and S. P. Godse. "Detection and Prevention of Phishing Websites Using Machine Learning Approach". *Proceedings of the 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, 2018, doi: 10.1109/ICCUBEA.2018.8697412.

[21] N. A. G. Arachchilage and S. Love. "A Game Design Framework for Avoiding Phishing Attacks". *Comput. Human Behav.*, 29(3), pages 706–714, 2013, doi: 10.1016/J.CHB.2012.12.018.

[22] N. A. G. Arachchilage and S. Love. "Security Awareness of Computer Users: A Phishing Threat Avoidance Perspective". *Comput. Human Behav.*, vol. 38, pages 304–312, 2014,

doi: 10.1016/J.CHB.2014.05.046.

[23] N. A. G. Arachchilage and M. Cole. "Designing a Mobile Game for Home Computer Users to Protect Against Phishing Attacks". *arXiv preprint arXiv:1602.03929*, 2016

[24] N. A. G. Arachchilage and S. Love. "A game design framework for avoiding phishing attacks". *Comput. Human Behav.*, 29(3), pages 706–714, 2013, doi: 10.1016/j.chb.2012.12.018.

[25] S.-S. Tseng, C.-H. Ku, T.-J. Lee, G.-G. Geng, and Y.-J. Wang. "Building a Frame-Based Anti-Phishing Model based on Phishing Ontology". *Proceedings of International Conference on Advances in Information Technology*, 2013.

[26] M. Bazarganigilani. "Phishing E-Mail Detection Using Ontology Concept and Naïve Bayes Algorithm". *Int. J. Res. Rev. Comput. Sci.*, 2(2), 2011.

[27] M. S. Qaseem and A. Govardhan. "Phishing Detection in IMs using Domain Ontology and CBA - An innovative Rule Generation Approach". *ArXiv preprint arXiv:1412.3056*, 2014.

[28] K. Kerremans, Y. Tang, R. Temmerman, and G. Zhao. "Towards Ontology-based E-mail Fraud Detection". *Proceedings of the 2005 Purtuguese Conference on Artificial Intelligence*, 2005, pages 106–111, doi: 10.1109/EPIA.2005.341275.

[29] G. Park. "Towards Ontology-Based Phishing Detection". Purdue University, 2018.

[30] Vamsee Krishna Kiran Muppavarapu, Ramesh Gowtham, and Archanaa Rajendran. "An RDF based Anti-Phishing Framework". *Int. Assoc. Sci. Innov. Res.*, 1(9), pages 1–10, 2013.

[31] C. Falk. "Knowledge Modeling of Phishing Emails". *Open Access Diss.*, Aug. 2016.

[32] J. Zhang, Q. Li, Q. Wang, T. Geng, X. Ouyang, and Y. Xin. "Parsing and Detecting Phishing Pages Based on Semantic Understanding of Text". *J. Inf. Comput. Sci.*, 9(6), pages 1521–1534, 2012.

[33] A. S. Bozkir and E. A. Sezer. "Use of HOG Descriptors in Phishing Detection". *Proceedings of the 2016 4th International Symposium on Digital Forensic and Security (ISDFS)*, 2016, pages 148–153, doi: 10.1109/ISDFS.2016.7473534.

[34] A. Oest, Y. Safaei, A. Doupé, G.-J. Ahn, B. Wardman, and K. Tyers. "PhishFarm: A Scalable Framework for Measuring the Effectiveness of Evasion Techniques Against Browser Phishing Blacklists". *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*, 2019, pages 764–781, doi: 10.1109/SP.2019.00049.

[35] N. Virvilis, A. Mylonas, N. Tsalis, and D. Gritzalis. "Security Busters: Web Browser Security vs. Rogue Sites". *Comput. Secur.*, vol. 52, pages 90–105, 2015, doi: 10.1016/J.COSE.2015.04.009.

[36] N. Tsalis, N. Virvilis, A. Mylonas, T. Apostolopoulos, and D. Gritzalis. "Browser Blacklists: The Utopia of Phishing Protection". *Springer*, pages 278–293, 2015.

[37] L. F. Sikos. "Description Logics: Formal Foundation for Web Ontology Engineering". in *Description Logics in Multimedia Reasoning*, Cham: Springer International Publishing, pages 67–120, 2017

[38] D. Ellison, A. R. Ikuesan, and H. Venter. "Description Logics and Axiom Formation for a Digital Forensics Ontology". *Proceedings of the European Conference on Cyber Warfare and Security*, pages 742–751, 2019

[39] N. Scarpato, N. D. Cilia, and M. Romano. "Reachability Matrix Ontology: A Cybersecurity Ontology". *Appl. Artif. Intell.*,

33(7), pages 643–655, 2019, doi: 10.1080/08839514.2019.1592344.

[40] G. Park and J. Rayz. "Ontological Detection of Phishing Emails". *Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2858–2863, 2018, doi: 10.1109/SMC.2018.00486.

[41] M. Benedek, Y. N. Kenett, K. Umdasch, D. Anaki, M. Faust, and A. C. Neubauer. "How semantic memory structure and intelligence contribute to creative thought: a network science approach". *Think. Reason.*, 23(2), pages 158–183, Apr. 2017, doi: 10.1080/13546783.2016.1278034.

[42] P. Di Maio and M. C. Suárez-Figueroa. "Introduction to the Special Issue 'Artificial Intelligence Knowledge Representation'". *Systems*, 7(3), p. 35, Jul. 2019, doi: 10.3390/systems7030035.

[43] A. Patel and S. Jain. "Formalisms of Representing Knowledge," in *Procedia Computer Science*, 2018, vol. 125, pages 542–549, doi: 10.1016/j.procs.2017.12.070.

[44] G. Jakus, V. Milutinović, S. Omerović, and S. Tomažič. "Concepts, Ontologies, and Knowledge Representation". *Springer*, 2013.

[45] V. Varga, C. Săcărea, and A. E. Molnar. "Conceptual Graphs Based Modeling of Semi-structured Data". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 10872 LNAI, pages 167–175, doi: 10.1007/978-3-319-91379-7_13.

[46] R. J. Brachman, "What's in a concept: structural foundations for semantic networks". *Int. J. Man. Mach. Stud.*, 9(2), pages 127–152, Mar. 1977, doi: 10.1016/S0020-7373(77)80017-5.

[47] R. Zakeri, R. Jalili, H. R. Shahriari, and H. Abolhassani, "Using Description Logics for Network Vulnerability Analysis". *Proceedings of International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies (ICNICONSMCL'06)*, pages 78–78, doi: 10.1109/ICNICONSMCL.2006.222.

[48] W. Yan, E. Hou, and N. Ansari. "Description logics for an autonomic IDS event analysis system". *Comput. Commun.*, 29(15), pages 2841–2852, 2006, doi: 10.1016/j.comcom.2005.10.038.

[49] T. Takahashi and Y. Kadobayashi. "Reference Ontology for Cybersecurity Operational Information". *Comput. J.*, 58(10), pages 2297–2312, 2015, doi: 10.1093/comjnl/bxu101.

[50] M. Krötzsch, F. Simančík, and I. Horrocks. "A Description Logic Primer *". 2013.

[51] F. Baader, I. Horrocks, C. Lutz, and U. Sattler. "*An Introduction to Description Logic*". Cambridge University Press, 2017.

[52] H. S. Shin. "Reasoning processes in clinical reasoning: from the perspective of cognitive psychology". *Korean J. Med. Educ.*, 31(4), pages 299–308, 2019, doi: 10.3946/kjme.2019.140.

[53] C. Lutz, U. Sattler, C. Tinelli, A.-Y. Turhan, and F. Wolter, Eds. "Description Logic, Theory Combination, and All That". *Springer International Publishing*, 2019.

[54] O. Curé and G. Blin. "Reasoning". *RDF Database Systems*, Morgan Kaufmann, 2015, pages 191–222.

[55] D. Allemang and J. Hendler. "*Semantic Web for the Working Ontologist*". Elsevier, 2011.

[56] C. Thomas. "*Ontology in Information Science*". InTech, 2018.

[57] K. Munir and M. Sheraz Anjum. "The use of

ontologies for effective knowledge modelling and information retrieval". *Applied Computing and Informatics*, 14(2), pages 116–126, 2018, doi: 10.1016/j.aci.2017.07.003.

[58] Z. Jin and Z. Jin. "Ontology-Oriented Interactive Environment Modeling". *Environ. Model. Requir. Eng. Softw. Intensive Syst.*, pages 45–67, 2018, doi: 10.1016/B978-0-12-801954-2.00004-2.

[59] M. A. Musen and the P. Protégé Team. "The Protégé Project: A Look Back and a Look Forward". *AI matters*, 1(4), pages 4–12, 2015, doi: 10.1145/2757001.2757003.

[60] R. Zese, E. Bellodi, F. Riguzzi, G. Cota, and E. Lamma. "Tableau reasoning for description logics and its extension to probabilities". *Ann. Math. Artif. Intell.*, 82(1–3), pages 101–130, 2018, doi: 10.1007/s10472-016-9529-3.

[61] G. Mohamed. "Raisonnement pour les Logiques de Description Appliqué Au Web Semantique". PhD thesis, Faculty of Mathematics and Computer Science, University of M'SILA, Algeria, 2016.