**Nom du projet** : Modular Automated Syntactic Signature Extraction
**Acronyme** : MASSE

**Porteur** : Teclib, 390 rue Saint-Honoré – 75001 Paris (PME)
**Consortium** : Teclib, Inria (équipe TAMIS, Rennes)

**Instrument de financement** : RAPID

**Dates** : du 1er septembre 2017 au 31 août 2019

**Description du projet :**

Le projet MASSE propose de développer une solution de production automatique de signatures YARA, utilisant les caractéristiques significatives du malware qui ont été utilisées pour sa classification. Ce mécanisme de production de signatures permettra de fournir des signatures plus efficaces dans la caractérisation des familles de malwares.

Les techniques d'analyse de malwares employées viseront des malwares difficiles à détecter par leur usage de polymorphisme ou de métamorphisme et présenteront une difficulté pour la rétro-ingénierie afin de gêner les créateurs de malware pour l'application de contre-mesures à la détection.

La solution proposée par le projet MASSE a deux caractéristiques essentielles :

- un traitement automatique des échantillons de malware, ne nécessitant pas l'intervention humaine pour l'analyse des codes malveillants et la production de signatures
- une architecture centralisée disposant d'une grosse puissance de calcul afin de mettre en œuvre des techniques d'analyse avancées et coûteuses et de fournir une réponse immédiate à tous les postes clients

L'architecture de MASSE est constituée :

- d'agents, fonctionnant sur Windows ou Linux, qui mettent en œuvre la détection temps réel à partir d'une base de règles YARA, l'envoi de fichiers suspects au 'cloud' d'analyse et la mise à jour de la base de règles
- d'un 'cloud' d'analyse assurant la classification des fichiers envoyés par les agents et, dans le cas où un fichier est classé comme malware, la génération automatique de règles YARA pour la détection

L'architecture d'analyse et de génération est modulaire et permet d'inclure facilement de nouveaux algorithmes de classification et de génération de règles.

La piste envisagée pour la génération de règles est basée sur les n-grams : l'algorithme extrait les séquences d'octets les plus significatives pour la discrimination malware/goodware, puis optimise les séquences obtenues à l'aide d'algorithmes génétiques.

Une architecture virtualisée de test et d'intégration continue est en cours de finalisation, permettant de tester plusieurs algorithmes de génération et de comparer leurs performances.