Vol.2011-CE-112 No.10
2011/12/18

# マルウェア対策研究人材育成ワークショップ ～教育コミュニティへの貢献とその課題～

寺田真敏

マルウェアによる脅威が複雑化する中，様々な対策研究が盛んに行われている．客観的な評価と研究成果の共有を容易にするため，サイバークリーンセンターで収集しているデータをもとに研究用データセット MWS Datasets(含む，CCC DATAset 2008-2010)を利用したマルウェア対策研究人材育成ワークショップ(MWS 2008-2011, http://www.iwsec.org/mws/2011/)を開催してきた．本稿では，マルウェア対策研究人材育成ワークショップ，研究用データセット，競技会などの活動について報告する．

# anti Malware engineering WorkShop (MWS) Contributions and challenges to the educational community

Masato Terada

There have been a lot of researches on countermeasures against the complicated threats by malware. anti-Malware engineering WorkShop (MWS) were held annually (2008-2011) in order to evaluate the proposals objectively and share the research achievements by using MWS Datasets (includes CCC DATAset 2008 - 2010). This paper presents an overview of MWS activities, which include an annual MWS workshop, MWS datasets, MWS Cup and so on.

## 1. はじめに

防衛産業を狙ったウイルス攻撃や金融機関を装う不審メールなど，マルウェア[a]を利用した脅威が複雑化する中，様々なアプローチでの対策研究が行われている．その一方で，研究を行う上で多くの課題があり，その一つに「共通の教材がないこと」が挙げられる．ここでの教材とは，提案手法の評価に用いるマルウェアのサンプルや，感染活動に関わる通信データなどのことである．教材となるこのような研究用データは，これまで研究者らが独自にハニーポット[b]を設置して収集し，それぞれの解析手法や対策手法の妥当性を検証するために利用してきた．このため，同じテーマに取り組む研究者同士であっても，研究成果を単純に比較することが難しい．また，新たに研究を始めようとしても，昨今のマルウェアに起因する感染事件の発生や所属組織のポリシーによる制約から「研究用データを収集すること自体が難しくなっていること」も課題となってきている．

現在，共通の教材として利用できる研究用データとして，DARPA Intrusion Detection Evaluation Data Sets[1]がある．このデータセットは，侵入検知システムの評価用であり，2000 年に公開されたデータが最新のものとなっている．このため，2001 年から 2003 年にかけインターネットで猛威を奮ったワームと呼ばれるマルウェアや，2004 年頃から出現したボットと呼ばれるマルウェアなど，新たな攻撃手法に関する素材が含まれてはいない．また，2009 年以降，サイバー防御演習時のデータセットである the 2009 Inter-Service Academy Cyber Defense Exercise datasets[2]，大規模セキュリティ関連データの収集と分析をもとに，より良いデータとナレッジの共有を図る BADGERS2011[3]，コンピュータ・ネットワークの運用データをレポジトリとして蓄積し，インフラ防護と脅威評価に活用する PREDICT[4]などのプロジェクトも動き始めてはいるが，マルウェアによる攻撃を想定したデータが含まれているわけではない．

このような課題を抱えている状況において，更なる進化を続けるマルウェアに対峙していくために，自分達に何ができるだろうか？，サイバークリーンセンターで収集しているデータを有効に活用できないだろうか？などを有志と語る中で，2008 年，研究用データセット(MWS Datasets)を研究者に提供し，研究成果を共有・切磋琢磨する場として「マルウェア対策研究人材育成ワークショップ(MWS)」を開催するに至った．

本稿では，教育コミュニティへの貢献とその課題の事例のひとつとして，2008 年から開催しているマルウェア対策研究人材育成ワークショップについて報告する．

a) マルウェアとは英語の Malicious (悪意のある) と Software を組み合わせた混成語であり，ユーザの望まない不正な動作を行うプログラムの総称である．
b) 感染手法などマルウェアに関わる多くの情報を得るためのおとり PC のことであり，マルウェアへの感染 PC を装う仕組みの総称である．

(株)日立製作所
Hitachi Ltd.

Vol.2011-CE-112 No.10
2011/12/18

## 2. マルウェア対策研究人材育成ワークショップ

### 2.1 開催の目的

インターネットのサイバー攻撃の脅威と実態など全般が見えにくくなってきている．その背景のひとつに，活動を見えにくくするためのマルウェア自身の機能高度化や運用が挙げられる．このような状況下で，情報システムでのセキュリティ事故や事件が発生した際に迅速に対処するためには，先端的な研究者だけではなく，企業のネットワーク技術ならびにセキュリティ技術を開発する実務者もマルウェアに関する専門知識を備えていく必要がある．そこで，本ワークショップでは，研究用データセットの提供，研究成果の共有ならびに切磋琢磨する環境の提供を通して，マルウェアに関する専門知識を備えた研究者／実務者を育成していくことを目的とした．

(1) 研究用データセットの提供

トラヒック分析技術やマルウェア分析技術を研究／評価するための適切な素材を準備し，研究者(学生，ネットワーク技術ならびにセキュリティ技術を開発する実務者)に提供することで，次の二点を実現する．

(2) 研究成果の共有

同じ研究用データセットを用いて行った研究成果を本ワークショップで発表し，研究者間で共有することで，より具体的な成果の水平展開を図り，我が国のセキュリティ研究人材育成につなげる．

(3) 切磋琢磨する環境の提供

同じ研究用データセットに基づく研究内容を共有することで，具体的なスキルアップ目標や，先進的な研究テーマの発見など，研究者の評価育成の場を形成する．

### 2.2 研究用データセットの提供から研究成果の共有の実現

マルウェア対策研究人材育成ワークショップの実現にあたっては，その活動を，研究用データセットの提供，分析ならびに対策技術の研究，研究成果の共有の 3 段階に分け，該当する既存コミュニティの協力を得るアプローチとした．具体的には，2008 年開始当初，「研究用データセット」についてはサイバークリーンセンター(https://www.ccc.go.jp/)からボット観測データの提供を受け，「研究成果の共有」「切磋琢磨する環境」の場として，情報処理学会コンピュータセキュリティ研究会で開催するコンピュータセキュリティシンポジウム(CSS)との併催とした(図 1).

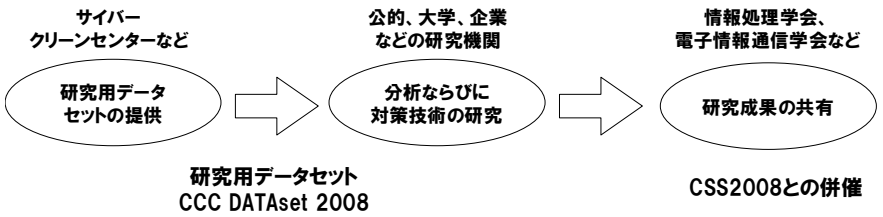


図 1：研究用データセットの提供から研究成果の共有

#### 2.2.1 研究用データセット

マルウェア対策研究人材育成ワークショップに提供している研究用データセット(MWS Datasets)を表 1，図 2 に示す[5][6][7].

表 1：MWS Datasets が提供する素材

| | | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|
| CCC DATAset | マルウェア検体 | ○ | ○ | ○ | ○ |
| | 攻撃通信データ | ○ | ○ | ○ | ○ |
| | 攻撃元データ | ○ | ○ | ○ | ○ |
| MARS | | | | ○ | |
| D3M | | | | ○ | ○ |

(1) CCC DATAset(Cyber Clean Center DATAset)

CCC DATAset は，サイバークリーンセンターが収集しているマルウェアの観測データ群である．検体解析技術の研究を想定した「マルウェア検体」，感染手法の検知ならびに解析技術の研究を想定した「攻撃通信データ」，活動傾向把握技術の研究を想定した「攻撃元データ」で構成される．

- マルウェア検体
  研究用データセットを提供するための観測装置(一般的に，おとり PC，ハニーポットと呼ばれている)で取得したマルウェアのハッシュ値
- 攻撃通信データ
  研究用データセットを提供するための観測装置で取得した通信のフルキャプチャデータ
- 攻撃元データ
  研究用データセットを提供するための観測装置で取得したマルウェア取得時のログデータ(マルウェア検体の取得時刻，送信元 IP アドレス，送信元ポート番号，

宛先 IP アドレス，宛先ポート番号，プロトコル，マルウェア検体のハッシュ値(SHA1)，ウイルス名称，ファイル名)

(2) MARS(Malware Analysis Result Set)

MARS は，検体となるプログラムを NICT(独立行政法人 情報通信研究機構)が所有する小規模攻撃再現テストベッドで実行した際に得られる動作記録データである．MWS Datasets では，検体となるプログラムとして CCC DATAset のマルウェア検体を使用している．

(3) D3M(Drive-by-Download Data by Marionette)

D3M は，NTT 情報流通プラットフォーム研究所が収集している Web 感染型マルウェアの観測データ群である．検体解析技術の研究を想定した「マルウェア検体」，感染手法の検知ならびに解析技術の研究を想定した「攻撃通信データ」で構成される．

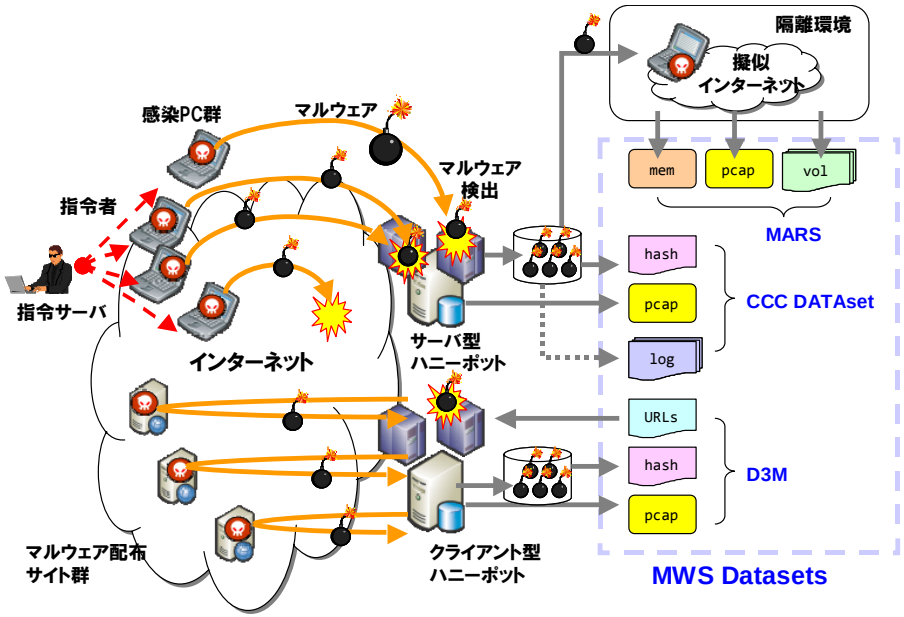


図 2：MWS Datasets を構成するデータ群

### 2.2.2 分析ならびに対策技術の研究

研究用データセットを入手した組織は，毎年 25 組織前後であり，半数以上が学術系となっている(表 2).

表 2：MWS Datasets の利用申請組織数

| | | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|
| 学術系 | 大学・大学院 | 18 | 17 | 17 | 15 |
| | 高専 | | | 1 | 1 |
| 企業系 | | 11 | 9 | 11 | 7 |
| 公的機関 | | 3 | 1 | | |
| 計 | | 32 | 27 | 29 | 23 |

*同一組織からの複数の利用申請は 1 件と算出

### 2.2.3 研究成果の共有

研究用データセットの概説，解析データ解説も含めて，毎年 20 件近くの研究発表が行われ，約半数は学生による発表となっている(表 3). 発表内容は，マルウェアの解析，検知，分類，感染防護に関する技術開発系の研究と，マルウェアの機能や活動に関する動向調査系の研究がある(図 3).

表 3：研究用データセット毎の発表件数

| | | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|
| CCC DATAset | マルウェア検体 | 5 | 10 | 6 | 6 |
| | 攻撃通信データ | 9 | 13 | 5 | 5 |
| | 攻撃元データ | 8 | 6 | 5 | 4 |
| MARS | | | | 1 | 1 |
| D3M | | | | 4 | 3 |
| その他 | | | 1 | 1 | 2 |
| 合計 | | 22(8) | 30(15) | 22(10) | 21(9) |

*括弧内は学生発表の件数

Vol.2011-CE-112 No.10
2011/12/18

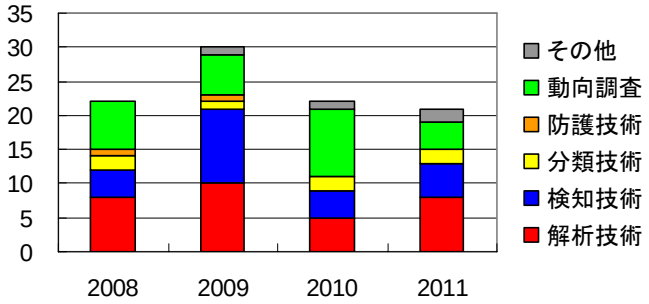


図 3：発表内容の分類

### 2.3 研究用データセットを用いた解析コンテスト MWS Cup

2009 年から「切磋琢磨する環境」の場を用意するため，日頃の研究・運用で培ったノウハウや開発ツールのレベルを競う MWS Cup を開始した．MWS Cup は，マルウェア対策に関する人材の育成と新たな技術やツールの発掘を目的とした，実践的な競技会である．競技課題に対する正答率を競うテクニカルコンポーネントと，解析に利用したノウハウや開発ツールの応用性をアピールするアーティステックコンポーネントから構成される．

#### (1) テクニカルコンポーネント

テクニカルコンポーネントの競技課題は，研究用データセット MWS Datasets を活用して, MWS Cup 企画委員会が作成している. 各年の競技課題は表 4 の通りである．競技時間は 90 分間で，正答率については，正解(True Positive：TP)のみを評価せず，運用の現場で問題となる誤検知(False Positive：FP)を減点対象とするところにも特徴がある．

#### (2) アーティステックコンポーネント

発表の部となるアーティステックコンポーネントでは，競技課題を解くために利用したノウハウや開発ツールが，教育や運用の現場で活用できうる美しさを持った方式であることを，3 分間でアピールして芸術点を競う．審査委員は，教育の現場で教材として活用でき得るかという「育成性」，産業の現場で監視ツールとして実用化でき得るかという「実用性」，審査委員の知見から総合的な美しさを評価する「芸術性」の 3 つの観点から採点する．

表 4：MWS Cup テクニカルコンポーネントの競技課題(1/2)

| 年 | 競技課題[c] |
|---|---|
| 2009 | 課題 1) 攻撃通信データを探し出せ<br>競技用 CD-ROM に含まれる 10 個の通信データファイルのうち，攻撃通信データが含まれる(B)と(B')の通信データファイルを探し出せ．<br>課題 2) マルウェア名を言え<br>上記で探し出した通信データファイルに関連するマルウェア名称を，CCC DATAset 2009 攻撃元データにある名称の中から述べ．ただし，名称は選択肢の記号から選ぶこと．<br>課題 3) 今後の通信パターンを予測せよ<br>上記で探し出したマルウェア通信を含む通信データファイルについて，今後の通信パターンを，選択肢の記号から選べ． |
| 2010 | 課題 1) 攻撃通信データを探し出せ<br>競技用 CD に収められている(W)が 35 個，(B)が 15 個の合計 50 個の通信データファイルから，(B)を探し出す．<br>課題 2) Web 感染を起こす悪性 Web サイトを特定せよ<br>競技用 CD に収められている，Web クライアントハニーポットの 5 個の通信データファイルを解析し，以下の各 URL を特定する．<br>● 入口 URL(2 問)<br>● 攻撃コードが含まれる Web サイトの URL(2 問)<br>● マルウェアを配布する Web サイトの URL(1 問) |
| 2011 | 課題 1) Web 感染を起こす悪性 Web サイトを特定せよ．<br>競技用 DVD に収められている 5 個の Web 閲覧通信データファイルについて以下の問いに答えよ．<br>第 1 問：攻撃コードが含まれるかどうかを答えよ．<br>第 2 問：攻撃コードが含まれるかどうかを答えよ．<br>第 3 問：入口 URL を答えよ．<br>第 4 問：攻撃コードが含まれる Web サイトの URL を答えよ．<br>第 5 問：マルウェアを配布する Web サイトの URL を答えよ． |

c) (W)はマルウェアに感染していないシステムの通信データ(正常通信), (B)はマルウェアが検出された部分を含む通信データ(マルウェア通信), (B')は正常通信(W)データとマルウェア通信(B)データを合成した通信データのこと．

表 4：MWS Cup テクニカルコンポーネントの競技課題(2/2)

| 年 | 競技課題 |
|---|---|
| 2011 | 課題 2) 攻撃の経緯を説明せよ．<br>どうやら VM(仮想計算機)上で動かしていた Windows がボットに感染したらしい．VM イメージとパケットキャプチャ情報を基に，以下の点を踏まえて攻撃の経緯を説明せよ．<br>(1)攻撃された脆弱性(CVE ナンバー) (選択肢より回答)<br>(2)脆弱性を突いて開放したバックドアポート (選択肢より回答)<br>(3)ボットをダウンロードし，実行した方法 (選択肢より回答)<br>(4)ボットの起動プロセス名<br>(5)ボットのファイルパス<br>課題 3) Android マルウェアを解析せよ．<br>5 個の Android アプリ(*.apk ファイル)の特徴を，以下の中から選択せよ．<br>(a) 正しいアプリ，(b) 情報漏えい型アプリ，(c) root 権限奪取型アプリ，(d) root 権限利用型アプリ，(e) root 権限奪取型＋端末改造型アプリ，(f) 情報漏えい型＋root 権限奪取型アプリ，(g) 勝手な SMS(C メール)送信型アプリ |

### 2.4 意見交換会と反省会

マルウェア対策研究人材育成ワークショップの特徴のひとつに，ワークショップ開催前の意見交換会と，終了後の反省会の開催が挙げられる．これら意見交換会と反省会は，研究用データセット利用者同士の顔合わせの場としてだけではなく，マルウェア対策研究人材育成ワークショップを発展させていくための場でもある．

意見交換会では，マルウェア対策研究人材育成ワークショップで使用する研究用データセットの概説や MWS Cup 開催要領と共に，契約手続きなど研究用データセットを利用するにあたっての諸注意を共有する．特に，マルウェア検体については，その検体を一意に特定するための情報となるハッシュ値を公開しない，検体をダウンロードできる悪意あるサイトの IP アドレスを開示しないなど，研究成果公開による副次的な影響を研究者自身が意識していく場としても活用している．

### 2.5 オープン化に向けて

2008 年～2010 年のマルウェア対策研究人材育成ワークショップ開催にあたっては，研究用データセット利用者を顔が見えている範囲をベースとした限定されたコミュニティとして運用してきた．この背景には，取り扱う研究対象が使い方によっては攻撃を助長する手段に変貌してしまう可能性があること，セキュリティ分野固有かもしれないが，信頼関係でつながることで(trust network)，潜在的に存在しうる脅威，特にユーザとして起こしてしまいやすいミスの低減を図ることにあった．さらに，2011 年 3 月，サイバークリーンセンターの運用が大きく変わったことで，マルウェア対策研究人材育成ワークショップにおける研究用データセットの提供段階の運用について，今まで以上に自立性が求められるようになってきた．

これまでの取り組みを継承し発展させていくためには，マルウェア対策研究人材育成ワークショップのオープン化，運営の定常化，コンピュータセキュリティシンポジウムとの融合の実現という課題を解決していく必要がある．そこで，2011 年から，情報処理学会コンピュータセキュリティ(CSEC)研究会の配下に，MWS 組織委員会を設置した(図 4).

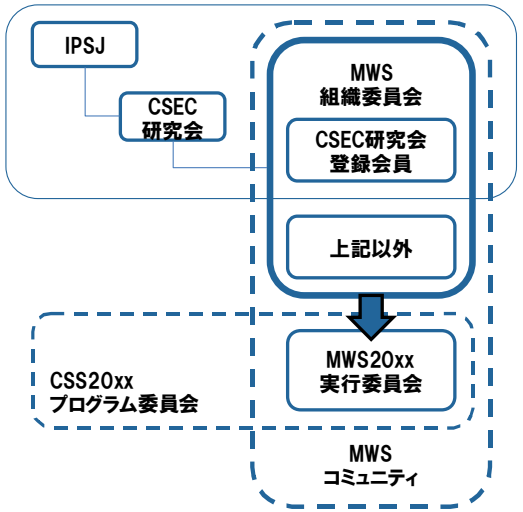


図 4：2011 年以降のマルウェア対策研究人材育成ワークショップの体制

商品名称等に関する表示
Windows は Microsoft Corporation の米国およびその他の国における登録商標または商標です．Android は，Google Inc の商標または登録商標です．本稿に記載されている会社名，製品名は，それぞれの会社の商標もしくは登録商標です．

MWS 組織委員会は，長期的な視点でこれら課題を解決しつつ，MWS の全体的な運営を担当する．MWS2011 実行委員会は，その年度のワークショップ開催の運営を担当する体制となっている．また，この体制では，学会活動という領域を維持しつつ，情報処理学会会員以外の関係者が，この活動に参加しやすい仕組みを整備するため，MWS コミュニティ，すなわち，マルウェア対策研究人材育成ワークショップという活動を主体とした仲間作りができるよう配慮している．

## 3. おわりに

本稿では，教育コミュニティへの貢献とその課題の事例のひとつとして，2008 年から開催しているマルウェア対策研究人材育成ワークショップについて報告した．マルウェア対策研究人材育成ワークショップは，進化を続けるマルウェアに対峙していくために，自分達に何ができるだろうか？という検討の中から生まれてきた活動であり，研究用データセット(MWS Datasets)を研究者に提供し，研究成果を共有・切磋琢磨する場として動き始めた．これらのマルウェア対策研究人材育成ワークショップの取り組みは，産官学のマルウェア対策に関わる実践的な取り組みとして，少しずつ認知されてつつあるが，まだ始まったばかりで，長期的な視点での運用上の課題はまだまだ残っている状況にある．また，研究用データセットの提供することだけではなく，研究用標準データ[d]自身が研究対象分野として立ち上がることを後押しする活動に発展させていきたいと考えている(図 5).

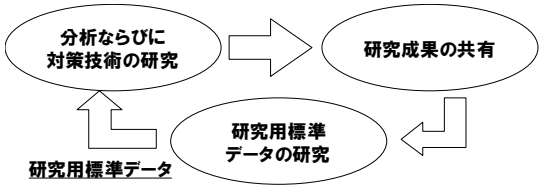


図 5：研究用標準データに基づくマルウェア対策研究サイクル

d) 研究用標準データの代表例として，1999 年に米リンカーン研究所が開発した"1999 DARPA Intrusion Detection Evaluation Data Set"がある．このデータは，侵入検知システムの有効性を確認するためのトラフィック評価データで，侵入検知技術の客観的な評価を行なうための評価データとしても活用されている．このような評価データは，技術の有効性や効果を客観的に確認するためのデータとして必要とされている．

## 謝辞

本活動を推進するにあたって，有益な助言とデータセット作成の協力を頂いたサイバークリーンセンターの関係者各位に深く感謝致します．

## 参考文献

1) MIT Lincoln Laboratory, DARPA Intrusion Detection Evaluation Data Sets, http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/index.html
2) B. Sangster, et al.: Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets, 18th USENIX Security Symposium CSET'09 (2009.08)
3) BADGERS2011: Building Analysis Datasets and Gathering Experience Returns for Security, http://iseclab.org/badgers2011/ (2011.04)
4) PREDICT: the Protected Repository for the Defense of Infrastructure Against Cyber Threats, https://www.predict.org/
5) 畑田充弘, 他: マルウェア対策のための研究用データセットとワークショップを通じた研究成果の共有, 情報処理学会シンポジウムシリーズ, Vol.2009, No.11, CSS2009(MWS2009), pp.1-8 (Oct. 2009)
6) 畑田充弘, 他: マルウェア対策のための研究用データセット ～MWS 2010 Datasets～, 情報処理学会シンポジウムシリーズ, Vol.2010, No.9, CSS2010(MWS2010), (Oct. 2010)
7) 畑田充弘, 他: マルウェア対策のための研究用データセット ～ MWS 2011 Datasets ～, 情報処理学会シンポジウムシリーズ, Vol.2011, CSS2011(MWS2011), (Oct. 2011)