# マルウェア対策のための研究用データセット ～ MWS 2010 Datasets ～

畑田 充弘[†1]　中津留 勇[†2]　秋山 満昭[†3]　三輪 信介[†4]

[†1] NTT コミュニケーションズ株式会社
〒108-8118 東京都港区芝浦 3-4-1 グランパークタワー17F
[†2] 一般社団法人 JPCERT コーディネーションセンター
〒101-0054 東京都千代田区神田錦町 3-17 廣瀬ビル 11F
[†3] NTT 情報流通プラットフォーム研究所
〒108-8585 東京都武蔵野市緑町 3-9-11
[†4] 独立行政法人 情報通信研究機構
〒184-8795 東京都小金井市貫井北町 4-2-1

E-mail: [†1] m.hatada@ntt.com, [†2] office@jpcert.or.jp, [†3] akiyama.mitsuaki@lab.ntt.co.jp, [†4] danna@nict.go.jp

**あらまし** 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(CCC DATAset 2008/2009)を利用したワークショップ(MWS2008/2009)を開催してきた．本稿では，MWS 2010 で利用する研究用データセット(MWS2010 Datasets)を構成する CCC DATAset 2010，マルウェア検体の動作記録データ(MARS)，Web 感染型マルウェアのデータセット(D3M 2010)の概要を報告する．

# Datasets for Anti-Malware Research ～ MWS 2010 Datasets ～

Mitsuhiro Hatada[†1]　You Nakatsuru[†2]　Mitsuaki Akiyama[†3]　Shinsuke Miwa[†4]

[†1] NTT Communications Corporation
Gran Park Tower 17F, 3-4-1 Shibaura, Minato-ku, Tokyo 108-8118, Japan
[†2] Japan Computer Emergency Response Team Coordination Center
3-17 Kandanishikicho, Chiyoda-ku, Tokyo 101-0054, Japan
[†3] NTT Information Sharing Platform Laboratories
Midori-Cho 3-9-11, Musashino, Tokyo 180-8585, Japan
[†4] National Institute of Information and Communications Technology
4-2-1 Nukui-Kita-Machi, Koganei-shi, Tokyo 184-8795, Japan

E-mail: [†1] m.hatada@ntt.com, [†2] office@jpcert.or.jp, [†3] akiyama.mitsuaki@lab.ntt.co.jp, [†4] danna@jaist.ac.jp

**Abstract** There has been a lot of researches on countermeasures against the complicated threats by malware. MWS 2008 and MWS 2009 were held in order to evaluate the proposals objectively and share the research achievements by using CCC DATAset 2008 and CCC DATAset 2009. This paper presents an overview of MWS 2010 Datasets for MWS 2010: CCC DATAset 2010, MARS, and D3M 2010.

## 1. はじめに

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．一方で，研究を行う上で様々な課題があり，その一つとして「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，感染前後の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．そのため，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．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

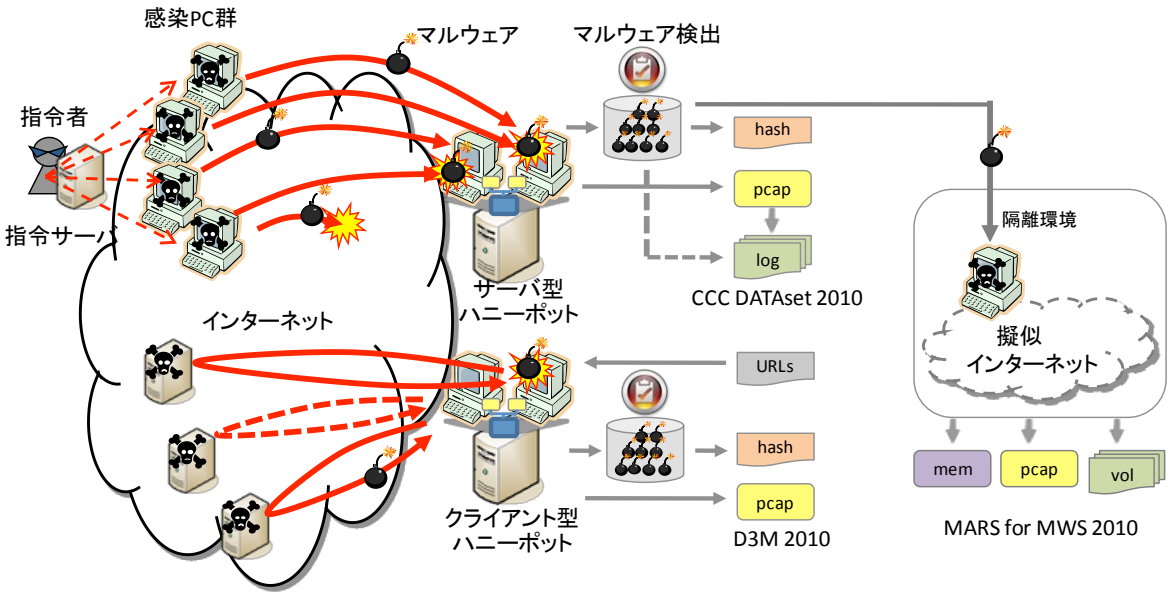


図 1 MWS 2010 Datasets の概要

インシデント事例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」も大きな課題である.

現在でも侵入検知システムの評価に用いられる DARPA Intrusion Detection Evaluation Data Sets[1]は, 最新のもので 2000 年のデータセットが公開されている. しかし, 2001 年の Code Red や Nimda, 2003 年の Slammer などインターネットで猛威を奮ったワームの出現, 2004 年頃から現在に至るボットネットによる脅威などへ大きく変化した攻撃手法が含まれていない. 近年のものでは the 2009 Inter-Service Academy Cyber Defense Exercise datasets[2]というサイバー防御演習時のデータセットが公開されているが, マルウェアによる攻撃を想定したものではない.

このような課題がある中で, 更なる進化を続けるマルウェアに対峙していくために, サイバークリーンセンター(CCC)[3]で収集しているデータを活用した研究用データセット:CCC DATAset 2008[4]を研究者に提供して, 研究成果を共有する場・切磋琢磨する環境として「マルウェア対策研究人材育成ワークショップ 2008(MWS2008)」を開催した. 22 件の発表(うち学生発表:8 件)と 2 件のパネルディスカッションを通して, 大学や研究機関に限らず産業界も交えた活発な議論を行った. CCC DATAset 2009[5]を提供した MWS2009 では, 28 件の発表(うち学生発表:15 件)と 1 件のパネルディスカッション, 研究用データセットを用いた新たな取り組みとして MWS Cup 2009[6]が行われた.

MWS2010[7]で利用する研究用データセット(MWS 2010 Datasets)は CCC DATAset 2010, マルウェア検体の動作記録データである MARS for MWS 2010(Malware Analysis Result Set for MWS 2010), Web 感染型マルウェアのデータセットである D3M 2010 ( Drive-by-Download Data by Marionette 2010)から構成され, 以下, 各章で概要を述べる.

## 2. CCC DATAset 2010

マルウェアの解析技術の研究のための「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のための「攻撃通信データ」, ボットの活動傾向把握技術の研究のための「攻撃元データ」の三つから構成される. 以下, それぞれについて概要を述べる .

### 2.1. マルウェア検体

ハニーポットで収集したマルウェア検体のハッシュ値(MD5, SHA1)50 個をテキスト形式で記載したファイルであり, 以下の観点で選定している.

(1) 解析結果を照合できる検体:10 検体

(2) 未知検体:40 検体

(1)は特徴的な機能を有し, 技術的に目を通しておきたい検体であり, 事前に静的解析が完了している. そのため, 解析精度の評価に活用するこ

とを考慮した要件である. 具体的には, ユーザの特定の動作をトリガとして動作する検体や, 独自かつ高度な通信プロトコルを使用する検体である. (2)は 2010 年 1 月から 3 月までに収集した未知検体のうち, 収集日が偏らないよう任意で選定した検体であり, 相当数の検体の自動解析や自動分類を考慮した要件である.

### 2.2. 攻撃通信データ

ハニーポットの通信を tcpdump でパケットキャプチャした libpcap 形式のファイルである. ハニーポットは, ホスト OS 上の 2 台(honey001, honey002)のゲスト OS がそれぞれインターネット接続されており, パケットキャプチャはホスト OS 上で行っている. ゲスト OS は, 2 台とも Windows XP SP1 であり, ゲスト OS は定期的にクリーンな状態にリセットされる. データ収集日は 2010 年 3 月 5 日から 3 月 11 日, 総パケット数が 22,486,674 パケット, 約 3.5GB のデータサイズである.

### 2.3. 攻撃元データ

2009 年 5 月 1 日から 2010 年 4 月 30 日までの 1 年間にハニーポットで記録したマルウェア取得時のログで, 表 1 に示す項目を 1 レコードとして記録した csv 形式のファイルである. Windows2000 が稼働するハニーポットも一部含み, 国内の複数の ISP にそれぞれ接続された 92 台のハニーポットで記録された約 156MB のデータである. 攻撃元データの基本情報を表 2 に示す.

マルウェア検体のダウンロードを開始した時刻がマルウェア検体の取得時刻であり, ゲスト OS の Windows 上でのファイル作成日時となる. 送信元 IP アドレスまたは宛先 IP アドレスにおいて, ハニーポットの IP アドレスは各ハニーポットに対応する ID(honey001～honey092)に置換されて記載されている. ウイルス名称は収集日の翌日午前 3 時の最新パターンファイルを適用したウイルススキャナ(トレンドマイクロ社製)により判定された名称であり, マルウェアとして判定されなかったものは UNKNOWN と表記される. このため, パターンファイルのウイルス名称が更新された場合, 同一のハッシュ値であっても, 異なるウイルス名称が付与される場合がある.

MWS2010 では, 過去のデータとの傾向を比較分析することができるよう CCC DATAset 2008 と CCC DATAset 2009 も参考情報として提供しており, それらの差異を表 3 にまとめる.

表 1 攻撃元データのログ項目と例

| ログ項目 | 例(一部を*でマスク) |
|---|---|
| マルウェア検体の取得時刻 | 2010-03-05 03:02:41 |
| 送信元 IP アドレス | honey001 |
| 送信元ポート番号 | 1028 |
| 宛先 IP アドレス | **. 243.167 |
| 宛先ポート番号 | 5824 |
| TCP または UDP | TCP |
| マルウェア検体のハッシュ値(SHA1) | *****bc3c850cf68a39c9e8013f2169d408a9d90 |
| ウイルス名称 | WORM_DOWNAD.AD |
| ファイル名 | C:¥WINDOWS¥system32¥dhnlr.dll |

表 2 攻撃元データの基本情報

| 項目 | 件数 |
|---|---|
| 全レコード数 | 1,162,093 |
| TCP によるダウンロードレコード数 | 1,053,977 |
| UDP によるダウンロードレコード数 | 108,116 |
| ダウンロードホスト IP アドレス種類数 | 176,522 |
| マルウェア検体のハッシュ値種類数 | 29,858 |
| ウイルス名称種類数(UNKNOWN 含まない) | 978 |

## 3. MARS for MWS 2010

研究用データセット MARS for MWS 2010 は, 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[8]で, CCC DATAset 2010 のマルウェア検体の検体実体を実際に動作させ, その観測結果などをまとめたデータセットである. 小規模攻撃再現テストベッドは, 閉鎖環境内で, 実機の上で実際の OS を使って, 実際のマルウェアの実行実体や小規模攻撃ツールを実行し, 用意された擬似インターネットへの通信やホスト内部での挙動を観測し, データセットとして出力する.

MWS2009 のパネルディスカッションでは, CCC DATAset 以外の研究用データセットの必要性について議論があり, その例として, マルウェア検体の解析を容易にする動作記録データを提供することとなった.

MARS for MWS 2010 は,「分類情報」「動的解析結果」「静的解析結果」の三つから構成される. 以下, それぞれについて概要を述べる.

表 3 CCC DATAset 2008／2009／2010 の差異比較

| 項目 | | 2008 | 2009 | 2010 |
|---|---|---|---|---|
| マルウェア検体 | | | | |
| | 検体数 | 1 | 10 | 50 |
| | 選定条件 | 多機能，解読困難 | 解析結果あり，関連性のある複数検体，特徴的な機能 | 解析結果あり，特徴的な機能，2010 年 1 月～3 月に収集した未知検体 |
| 攻撃通信データ | | | | |
| | ハニーポット | honey001，honey002 | honey003, honey004 | honey001，honey002 |
| | 収集日 | 2008/4/28，2008/4/29 | 2009/3/13，2009/3/14 | 2010/3/5～2010/3/11 |
| | 総パケット数 | 15,901,943 | 3,511,850 | 22,486,674 |
| 攻撃元データ | | | | |
| | ハニーポット数 | 112 台 | 94 台 | 92 台 |
| | ハニーポット ID | なし（ダウンロードホストと通信方向のみ） | あり | あり |
| | 収集期間 | 2007/11/1～2008/4/30 | 2008/5/1～2009/4/30 | 2009/5/1～2010/4/30 |
| | 全レコード数 | 2,942,221 | 2,470,766 | 1,162,093 |

### 3.1. 分類情報

検体実体のファイル名や各種ハッシュ値，簡易的な分類情報が含まれる検体実体のメタ情報と，解析に用いた検体実体への参照と各結果ファイルの情報が含まれる動作記録データのメタ情報であり，ともに XML 文書として提供する.

### 3.2. 動的解析結果

マルウェア検体実体動作時のメモリダンプやパケットダンプ，及び擬似インターネットの一部である模倣 DNS へのアクセス記録である. なお，メモリダンプは検体実体を動作させてから 1 分後の物理メモリをすべて取得しており仮想メモリは含まない. パケットダンプは検体実体を動作させる直前から 5 分間取得している.

### 3.3. 静的解析結果

検体実体の strings コマンド結果や，動的解析結果として取得したメモリダンプに対する Volatility Framework[9]による各種解析結果である.

また，参考情報として CCC DATAset 2008 及び CCC DATAset 2009 のマルウェア検体の動作記録データである MARS for MWS 2008 及び MARS for MWS 2009 も MWS2010 で提供する.

## 4. D3M 2010

D3M 2010 は，NTT 情報流通プラットフォーム研究所の高対話型の Web クライアントハニーポット(Marionette[10])で収集したマルウェア検体，攻撃通信データの 2 つを収録した Web 感染型マルウェアの観測データ群である.

Marionette は脆弱性に対する攻撃を受けるがダウンロードされたマルウェアの実行を許可しない. そのため，CCC DATAset の攻撃通信データとは異なり，感染後のマルウェアの通信挙動は D3M 2010 の攻撃通信データには含まれない.

CCC DATAset はいわゆるサーバ型ハニーポットで収集したデータであり，近年脅威となっている Web ブラウザの脆弱性を利用して制御を奪い，マルウェアを強制的にダウンロード及びインストールさせる Drive-by-download 攻撃を捉えた研究用データセットへの必要性から提供することとなった.

D3M 2010 は，マルウェアの解析技術の研究のための「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のための「攻撃通信データ」から構成される. 以下，それぞれについて概要を述べる .

### 4.1. マルウェア検体

Web クライアントハニーポットで収集した”Gumblar.8080 系”のマルウェア検体のハッシュ値(3体分)をテキスト形式で記載したファイルである. 2010 年 3 月 8 日～11 日に収集した検体であり, 攻撃通信データには含まれていない検体である.

### 4.2. 攻撃通信データ

Web クライアントハニーポット 10 台の通信を tcpdump でパケットキャプチャした libpcap 形式のファイルである. ハニーポットの OS は Windows XP SP2, ブラウザは Internet Explorer 6.0, プラグインが Adobe Reader, Flash Player, WinZip, QuickTime, JRE であり, 何れもセキュリティパッチは未適用である. 10 台それぞれがインターネット接続されており, パケットキャプチャは上流ネットワークにあるスイッチのミラーポートで行っている. データ収集日は 2010 年 3 月 8 日, 9 日, 11 日であり, 日毎に 1 ファイル, 計 3 ファイルで約 133MB である.

巡回対象 URL は公開ブラックリスト(malwaredomainlist.com)に登録されている URL の中から, 各データ収集日に攻撃を検知した URL を予め抽出したものを用い, 参考情報として提供している.

## 5. おわりに

近年活発に行われているマルウェア対策研究において, 研究素材あるいは客観的な評価のための研究用データセットとなりえる MWS 2010 Datasets について述べた.

研究用データセット自身が研究者間での共通言語としての役割を担うことや, 研究用データセットとともに研究に用いたツールや解析したデータが共有されれば人材育成を含む本研究分野の発展に寄与することが期待できる. 今後, 最新の脅威を捉えた研究用データセットの収集・作成と利用環境の構築・提供など包括的なフレームワークを検討するとともに, 評価用として利用可能な研究用標準データの作成に向け検討していきたい.

### 謝辞

本研究にあたって, 有益な助言とデータセット作成の協力を頂いた CCC 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 ) サイバークリーンセンター, https://www.ccc.go.jp/

4 ) 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 (2009.10)

5 ) 畑田充弘: 研究用データセットを用いたマルウェア対策研究人材育成ワークショップ, 情報処理, Vol.51, No.3, pp.284-287 (2010.03)

6 ) 竹森敬祐, 他: MWS Cup 2009, 情報処理, Vol.51, No.3, pp.296-299 (2010.03)

7 ) マルウェア対策研究人材育成ワークショップ, http://www.iwsec.org/mws/2010/

8 ) 三輪信介, 他: 小規模攻撃再現テストベッドによる動作記録データセットの生成, 情報処理学会シンポジウムシリーズ, Vol.2009, No.11, CSS2009(MWS2009), pp.931-936 (2009.10)

9 ) N. Petroni, et al: FATKit: A Framework for the Extractionand Analysis of Digital Forensic Data from Volatile System Memory, Digital Investigation Journal 3(4) (2006.12)

10) Mitsuaki Akiyama, et al: Design and Implementation of High Interaction Client Honeypot for Drive-by-download Attacks, IEICE Transactions on Communication, Vol.E93-B No.5 pp.1131-1139 (2010.05)