DIGITAL FINGERPRINTS: IMPLEMENTING ALGORITHMS AS TECHNICAL CONTROLS

Gregg Gazvoda, M.S.

Manager, Computer System Validation

Covance Inc.

03 June 2016

SQA Data Integrity Conference

Denver, CO

COVANCE
SOLUTIONS MADE REAL®

Overview

KEY TOPICS

- Problem with Evolution
- Data Collection Systems
- Procedural and Technical Controls
- Hash Function
- MD5 Algorithm
- How to Utilize
- Risks
- Summary
- References

Image, Covance

Problem with Evolution

PHYSICAL MEDIA VS. DIGITAL

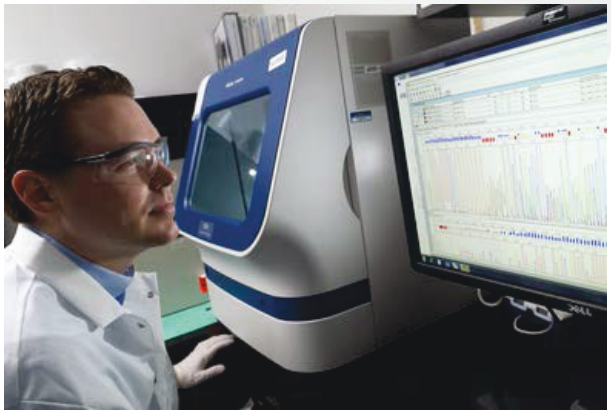
- Over the last 20 years we have seen a growing trend
  - Physical media is being replaced by digital
    - e.g., CDs vs. mp3
- Evolution is a good thing, right?
  - Easier to acquire
  - Easier to use
  - Compact
- So what is the issue?
  - Data Integrity

Image, Covance

Data Collection Systems

HOW WE COLLECT DATA HAS CHANGED

- Paper lab notebooks are a becoming a thing of the past
- Electronic systems used for capturing original observations
- The complexity and capability of these systems varies
  - Laboratory Information Management Systems (LIMS)
    - Provide the controls necessary to assure data integrity
      - System architecture (database)
      - Security
      - Audit trail, etc...
  - Standalone system (e.g., Instrument systems)
    - Often produce a single data file
    - Varying levels of security and attribution



Image, Covance

Data Collection Systems

HOW DO WE HANDLE THESE DATA FILES?

LIMS

- Under Control

Standalone

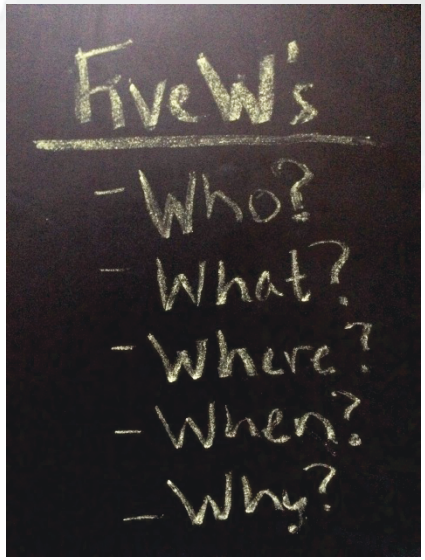
- Flat File: A flat file is a file containing records that have no structured interrelationship (Rouse, 2006)
  - e.g., Analyst
    - Produces a .wiff file
    - Stored locally or on a network
- In these cases, one needs to consider the integrity of the data files stored outside of a controlled system
- These systems may lack the necessary controls to detect or prevent modification
- Depending on the scope work being performed, this may require procedural and/or technical controls to claim compliance
    - GxP, 21 CFR Part 11, etc…

Procedural and Technical Controls

HOW DO WE ADDRESS THE 5 W'S (AND H)?

Who did what, when, where, why, and how?

- For systems lacking built-in controls, there may be a need for procedural or technical controls
  - Risk mitigation
- There are many solutions available on the market
  - e.g., Data sweeping tools, Document Management Systems, etc…
- These options may not always be practical
- Algorithms, often referred to as a hash function, hash sum, and/or checksum, provides an effective means of ensuring the integrity of data in this scenario



Image, Gazvoda

Hash Function

WHAT IS IT?

- A hash function/ hash sum/ checksum is a fixed-size datum computed from an arbitrary block of digital data for the purpose of detecting errors that may have been introduced during its transmission or storage
  - i.e. Digital Fingerprint
- Once a file is generated, a hash function can be generated that is unique to that file
  - The hash function of this presentation is:
    - MD5: A6F7CE2EEB33BD0CDE4565753D9A3E44
    - SHA-1: 734742FB2DF09C5E2C27D1222DCF25B7832D7AF9
- The integrity of the file can be checked at any later time by re-computing the hash function and comparing it with the original recorded at the time of creation (ICH, 2010)

Hash Function

HOW DOES IT WORK?

It's just 1s, 0s, and math!

- A bit of an understatement
- There are many types of hash functions to select from
- Some of the more common are:
  - MD5 (Message-Digest Algorithm)
  - SHA-1 (Secure Hash Algorithm 1)
  - SHA-256 (Secure Hash Algorithm 2 - 256-bit)
- Figure 1: the MD5 algorithm consists of 64 of these operations, grouped in four rounds of 16 operations

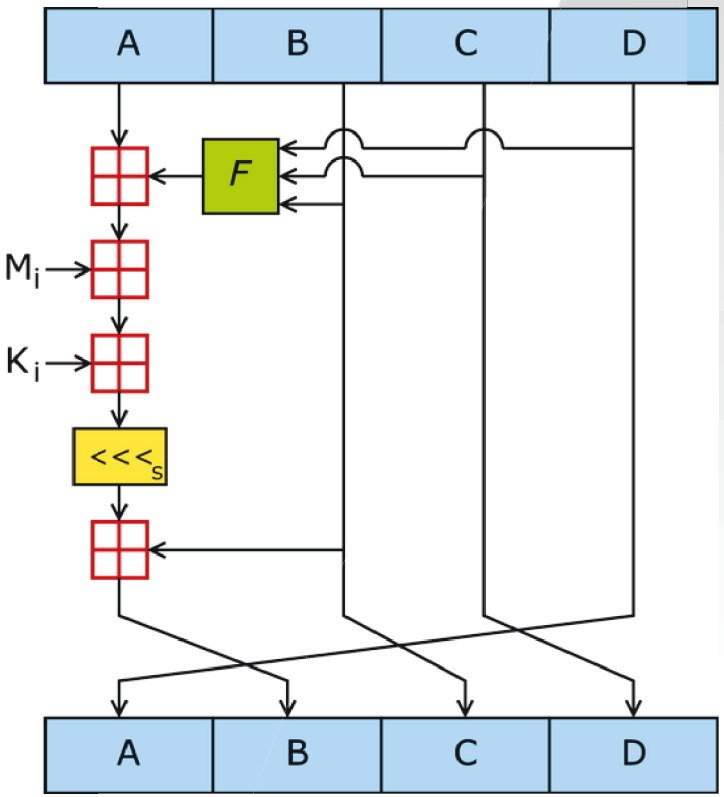


Figure 1: One MD5 operation (Crypto 2004)

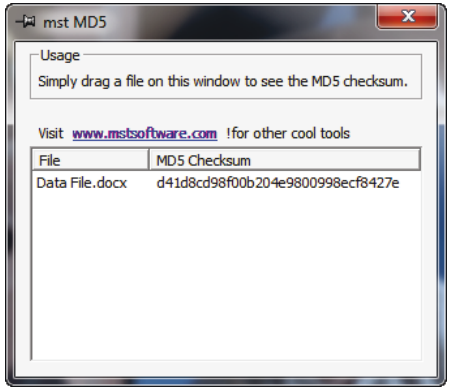
MD5 Algorithm

A DEEPER DIVE

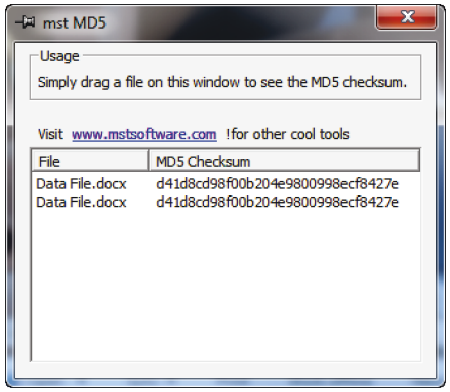
- The algorithm takes a data file of arbitrary length and produces, as output, a 128-bit, 32 digit hexadecimal "fingerprint" based on the input file (Rivest, 1992)
- This 32 digit hexadecimal number is unique to the file
- The checksum will remain unchanged as long as the data file itself is not modified
- It works with any file type of arbitrary length/size
- Multiple tools are readily available online
- Widely used across multiple industries

MD5 Algorithm

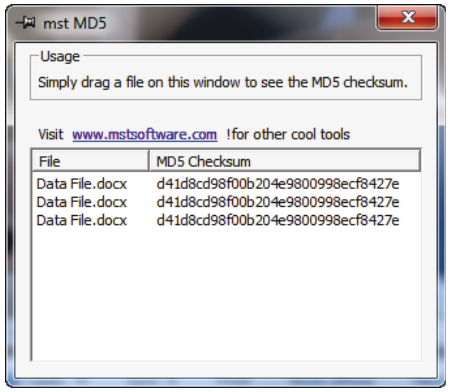
EXAMPLE CHECKSUM OF AN EXAMPLE DATA FILE:



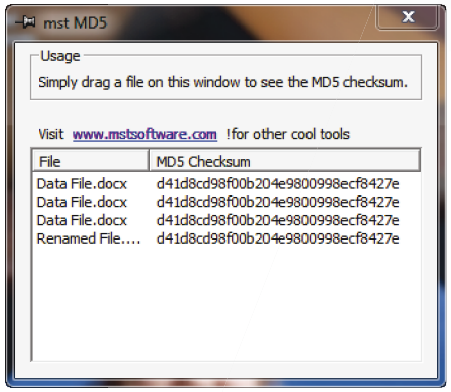
File moved to network



Open, changed, but not saved



File renamed



Data changed and saved

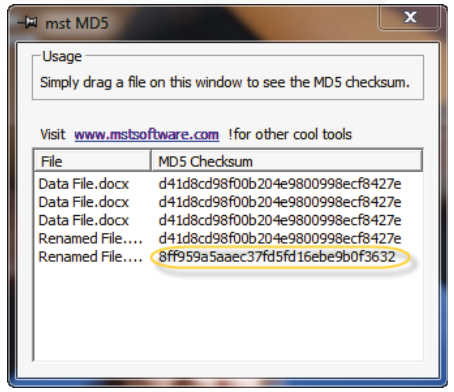


Figure 2: All images generated using the mst MD5 v2.0

How to Utilize

BUILD IT INTO YOUR SYSTEM/PROCESS

There are several options:

- ► Build these algorithms into your system(s)
  - Assuming you have the means necessary to implement this into your coding
- ► Adopt these algorithms as a technical/procedural controls
  - A more likely scenario

What does that look like?

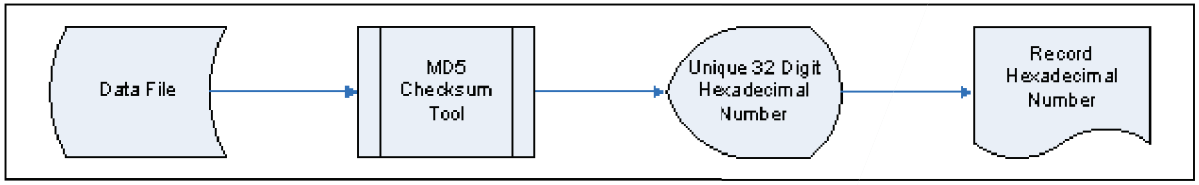


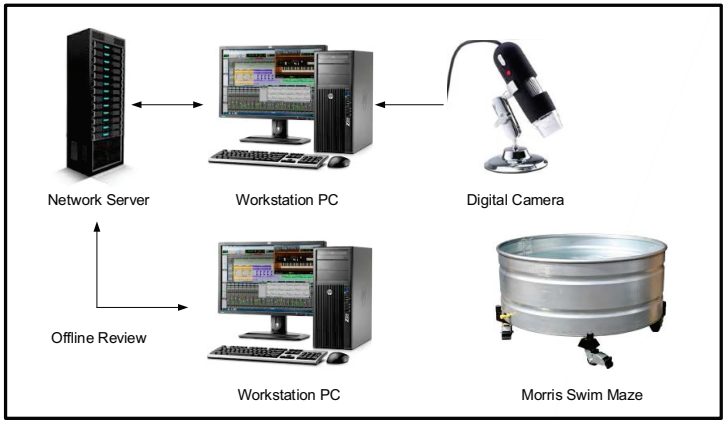
Figure 3: Proposed use of the MD5 Checksum tool - *Image, Covance*

How to Utilize

EXAMPLE:

Video capture system used to detect changes in spatial learning and memory in rodents

- The system produces a proprietary data file that is saved locally
- The system lacks the necessary controls to claim Part 11 compliance
  - No security
  - No audit trail
- As a result, need to address this as a hybrid system to ensure attribution of the collected data
  - How best to do so?

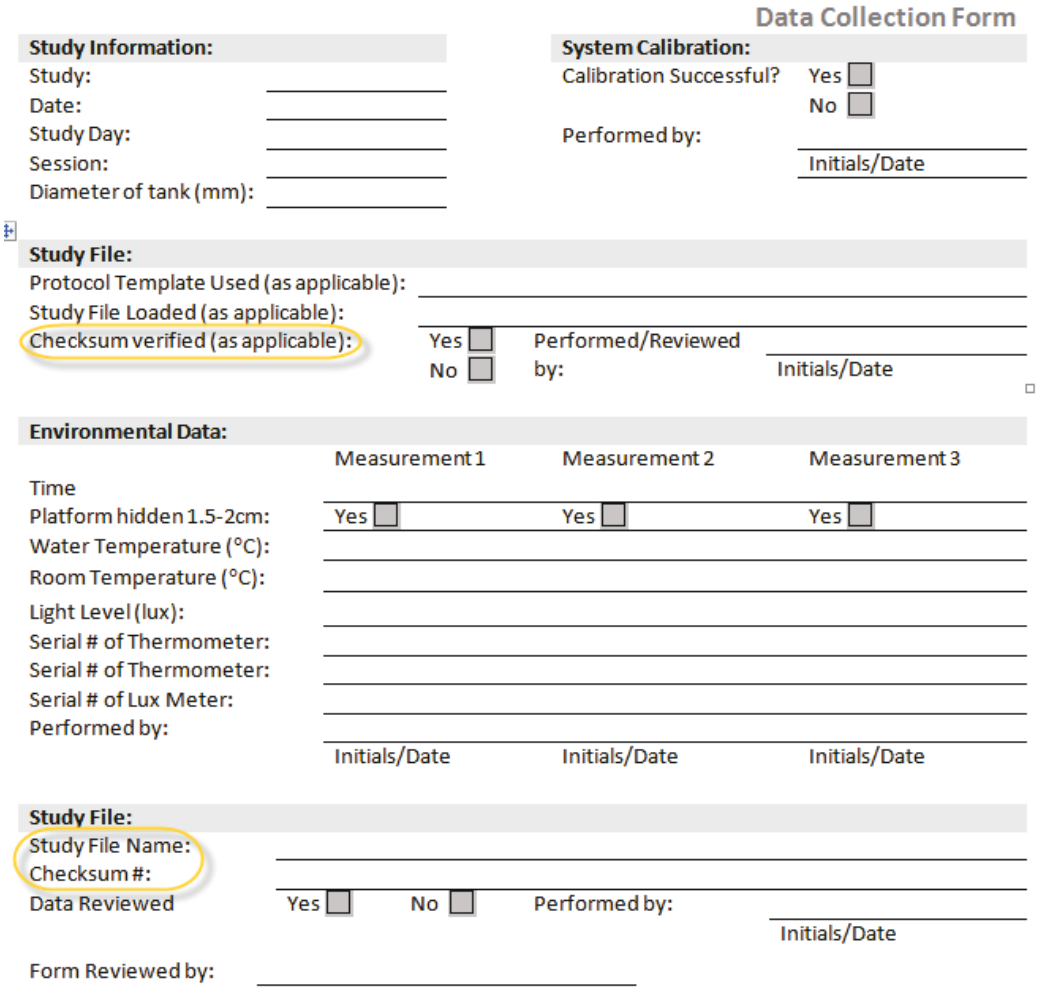


Video Capture System

Figure 4: Video Capture System - *Image, Covance*

How to Utilize

IF IT WASN'T DOCUMENTED…



Data Collection Form

Study Information:
Study: _____
Date: _____
Study Day: _____
Session: _____
Diameter of tank (mm): _____

System Calibration:
Calibration Successful? Yes ☐ No ☐
Performed by: _____ Initials/Date

Study File:
Protocol Template Used (as applicable): _____
Study File Loaded (as applicable): _____
Checksum verified (as applicable): Yes ☐ No ☐ Performed/Reviewed by: _____ Initials/Date

Environmental Data:

	Measurement 1	Measurement 2	Measurement 3
Time			
Platform hidden 1.5-2cm:	Yes ☐	Yes ☐	Yes ☐
Water Temperature (°C):			
Room Temperature (°C):			
Light Level (lux):			
Serial # of Thermometer:			
Serial # of Thermometer:			
Serial # of Lux Meter:			
Performed by:	Initials/Date	Initials/Date	Initials/Date

Study File:
Study File Name: _____
Checksum #: _____
Data Reviewed Yes ☐ No ☐ Performed by: _____ Initials/Date

Form Reviewed by: _____

Figure 5: Example Form - *Image, Covance*

COVANCE
SOLUTIONS MADE REAL®

How to Utilize

IF IT WASN'T DOCUMENTED…

Data Collection Form

Study Information:

Study:	1234-567
Date:	03 Jun2016
Study Day:	Day 1
Session:	1
Diameter of tank (mm):	1480

System Calibration:

Calibration Successful? Yes ☑ No ☐

Performed by: GG 03 Jun 2016 (Initials/Date)

Study File:

Protocol Template Used (as applicable): Morris Maze 1

Study File Loaded (as applicable): N/A

Checksum verified (as applicable): Yes ☑ No ☐ Performed/Reviewed by: GG 03 Jun 2016 (Initials/Date)

Environmental Data:

	Measurement 1	Measurement 2	Measurement 3
Time	0600	1200	1600
Platform hidden 1.5-2cm:	Yes ☑	Yes ☑	Yes ☑
Water Temperature (°C):	25	24	25
Room Temperature (°C):	25	25	25
Light Level (lux):	14	13	14
Serial # of Thermometer:	T12875	T12875	T12875
Serial # of Thermometer:	R231	R231	R231
Serial # of Lux Meter:	L45Q	L45Q	L45Q
Performed by:	GG 03 Jun 2016	GG 03 Jun 2016	SW 03 Jun2016
	Initials/Date	Initials/Date	Initials/Date

Study File:

Study File Name: study1234-567-day1.file

Checksum #: 77Ce03f75d6756e57a052C74e17b90C5

Data Reviewed Yes ☑ No ☐ Performed by: GG 03 Jun 2016 (Initials/Date)

Form Reviewed by: CG 08 Jun2016

Figure 5: Example Form - *Image, Covance*

COVANCE
SOLUTIONS MADE REAL®

Risks

THERE ARE SOME

As computational power has increased, we are beginning to see that some hash functions are susceptible to collision and/or a chosen-prefix collision attacks

- This was demonstrated for the MD5 algorithm and SHA-1 (Dobbertin 1996) (Stevens 2007, 2009)

As such, some hash functions may not be considered appropriate for SSL certificates

---

While these attacks call into question the security and reliability of these hash functions, they are not something that a typical user of ordinary means is capable of producing

- There are other factors which can help mitigate the risk

Summary

- Broad application
- Not perfect, but good enough
- Pros outweigh the Cons
- If implemented properly, can drastically improve the integrity of hybrid systems
- Validate (as necessary)

Questions

Image, Covance

References

- Rouse, M. (2006). *Flat File.* Retrieved from: http://searchsqlserver.techtarget.com/definition/flat-file (last accessed 27 May 2016)
- International Conference on Harmonisation (ICH). (2010). *EWG M2 Recommendation to the ICH Steering Committee Electronic Standards for the Transfer of Regulatory Information (ESTRI) File Integrity – MD5. Version 1.0.* Retrieved from http://www.ich.org/ (last access 11 June 2013)
- Crypto, M. (2004). *Figure 1*. Retrieved from: http://en.wikipedia.org/wiki/Image:MD5.png (last accessed 27 May 2016)
- Rivest, R.L. (1992). *The MD5 Message-Digest Algorithm.* RFC 1321, MIT Laboratory for Computer Science and RSA Data Security, Inc.,

References

- Dobbertin, H. (1996). *The Status of MD5 After a Recent Attack.* RSA Laboratories CryptoBytes 2 (2): 1. Retrieved 2013-06-06.
- Stevens, M., et al. (2009). *Chosen-prefix Collisions for MD5 and Applications*. Retrieved 2013-06-06.
- Stevens, M., et al. (2007) *Vulnerability of software integrity and code signing applications to chosen-prefix collisions for MD5*. Retrieved 2013-06-06.

Thank You

Covance Inc., headquartered in Princeton, NJ, is the drug development business of Laboratory Corporation of America® Holdings (LabCorp®). Covance is the marketing name for Covance Inc. and its subsidiaries around the world.

COVANCE
SOLUTIONS MADE REAL®