CROSS PLATFORM EMAIL DUPLICATE IDENTIFICATION

EDRM Duplicate Identification Project Team

11 January 2023

TABLE OF CONTENTS

DUPLICATE IDENTIFICATION PROJECT OVERVIEW

During discovery, disclosure or an investigation, it is often useful to identify duplicate emails in data exchanged between parties. This can deliver many benefits, including the ability for legal teams to rapidly triage emails already reviewed that also reside in data received from others.

While current approaches effectively identify email duplicates within native datasets processed by a single vendor, they do not enable duplicate identification across emails processed by multiple vendor platforms. Vendors use similar methods to detect email duplicates, but there are nuanced differences in their proprietary algorithms.

Currently no means of cross platform email duplicate identification exists, except to reprocess the data using a single vendor platform, often expending significant time and cost. The EDRM Duplicate Identification project set out to develop a solution to cross platform email duplicate identification.

Our solution is a simple, but effective approach which involves the use of the hash value of an email Message ID metadata field that we have named the *EDRM Message Identification Hash ("MIH").* This new approach need not replace current vendor email deduplication methods, but will enable cross platform email duplicate identification. It is expected cross platform duplicate identification using the MIH will be applied to email data sets that have already been deduplicated using a vendor's standard deduplication process. It is envisioned that the *EDRM MIH* will be an additional field that will be generated as part of the processing functionality in each vendor's platform and used by recipients to further identify duplicates for collections with established EDRM MIH values. This email duplicate identification process can be used for various purposes, including grouping email duplicates to enable a more efficient review process or deduplicating a cross platform email dataset.

The EDRM Email Duplicate Identification Specification Committee ("Committee") has developed the EDRM Email Duplicate Identification Toolkit ("Toolkit") to facilitate cross platform identification of duplicate email messages. The Committee anticipates that the use of EDRM MIH will lead to significant time and cost benefits.

The Toolkit is designed for a range of stakeholders:

- *Parties* who are encouraged to exchange EDRM MIH values as additional fields in their produced metadata to facilitate efficient identification of duplicate emails across data received from other parties irrespective of the vendor platform used to produce it.
- *Vendors* who are encouraged to add the calculation of EDRM MIH values to their processing and production toolsets and methodologies. They do not need to replace current code or change the way they deduplicate email messages internally within their platforms.
- *Regulators & Courts* who are encouraged to include the EDRM MIH in exchange protocols and practice notes.

This document includes two components of the Toolkit:

- ***EDRM Message Identification Hash (MIH) Specification (v1.0)*** is a succinct, technical specification with advisory notes geared to software developers and has been written for the target audience of vendors who are implementing the MIH in their platform. It defines a process to identify duplicate emails across disparate formats and forms of data employed in electronic discovery and disclosure.
- ***EDRM Email Duplicate Identification Guidelines (v1.0)*** is a non-technical reference for those who need to understand why and how to use the MIH. It outlines the objectives, methodology, potential use cases, advantages, and usage considerations of the Specification. These Guidelines are intended for use by those who want to use the MIH for cross platform duplicate identification, including parties and counsel, vendors and service providers and regulators and courts.

Other components of the Toolkit are:

- ***Whitepaper*** is a practical, non-technical introduction to the use of the EDRM MIH and is a useful tool for lawyers who need to quickly understand the benefits of using the MIH.
- ***Infographic*** is a simple one-page explanation of the solution.
- **Data and utilities to support use of the EDRM MIH Specification**
    - **Test Data Set -** a corpus of emails enabling testing and verification of EDRM MIH implementations by vendors.
    - **Small Dataset MIH Calculator –** an Excel-based tool to generate EDRM MIH values for small sets of Message-IDs.
    - **Open-source code with GUI frontend to extract MSGID from emails**, calculate EDRM MIH values from Outlook PST containers and output an email identifier, MSGID & MIH in a CSV file.

All components of the Toolset will be made accessible from the EDRM website.

CONTRIBUTORS

The project team (organizations noted for identification purposes only) includes:

- Murali Baddula, Chief Digital Officer at Law In Order (Sydney, Australia)
- Craig Ball, Attorney, Certified Computer Forensic Examiner and Adjunct Professor, University of Texas School of Law (Austin, TX USA)
- Ian Folkman, Vice President - Forensic Technology at Deloitte Tohmatsu Financial Advisory (Japan)
- Scott Foster, Senior Managing Director at FTI Consulting - Technology Lead (Australia)
- Matthew Golab, Director Legal Informatics and R+D at Gilbert + Tobin (Sydney, Australia)
- Phil Haselden, Technical Fellow at EDT (Australia)
- Greg Houston, Workflow Enablement at Relativity (Chicago, IL USA)
- Dr Paul Hunter, Chief Data Scientist at EDT (Australia)
- Dinesh Karamchandani, Senior Data Scientist at Reveal-Brainspace (Chicago, IL USA)
- Lisa Kozaris, Chief Innovation & Legal Solutions Officer at Allens (Melbourne, Australia)
- Enzo Lisciotto, Director of Customer Success at Reveal-Brainspace & Co-Membership Director of ACEDS ANZ (Sydney, Australia)
- James MacGregor, Partner at FORCYD (London, UK)
- Karan Mehta, Head of Legal Technology at Allens (Sydney, Australia)
- Rachi Messing, Co-Founder at Altorney (Israel)
- Elizabeth Miller, Senior Director at Law in Order (Australia)
- Beth Patterson, Director at ESPconnect & Adjunct Professor at University of Technology Sydney Law School (Australia)
- Alexander Poelma, Product Manager Nuix Workstation & Engine at Nuix (California, USA)
- Jo Sherman, CEO & Founder at EDT (Australia/USA)
- Paul Sirkis, Division Director, Regional Head of Litigation, Americas at Macquarie Group (New York, NY USA)
- George Socha, Senior Vice President of Brand Awareness at Reveal-Brainspace (Chicago, IL USA)
- Stephen Stewart, CTO at Nuix (Philadelphia, PA USA)
- Gavin Wingfield, Director Applied Legal Technology at King & Wood Mallesons
- Emma Young, Head of eDiscovery Delivery at Sky Discovery (Brisbane, Australia)

EDRM EMAIL DUPLICATE IDENTIFICATION SPECIFICATION (V1.0)

The EDRM Email Duplicate Identification Specification defines a methodology used to identify duplicate emails, including sets supplied in varying formats and forms of production.

The EDRM Message Identification Hash (MIH) is the MD5 hash value of the ASCII string comprised of the Message-ID header field of RFC-compliant email messages.

SPECIFICATION FOR COMPLIANCE

1. The EDRM Message Identification Hash (MIH) MUST NOT be generated if an email has no Message-ID or the email has an invalid Message-ID value. A description of what constitutes a valid Message-ID value follows.
2. The complete Message-ID value, including the flanking angle brackets, MUST be used to generate the MIH.
3. The character case of the Message-ID value MUST NOT be altered before MIH generation. Changing case changes the hash value of the string.
4. Where more than one Message-ID value is contained within an email, the MIH MUST be generated using only the first Message-ID value declared in the parent email message headers.
5. The MIH MUST NOT be generated for items which are not email messages. This document does not cover other types of items which may have Message-IDs such as calendar or contact items.

VALID MESSAGE-ID VALUES

For the purposes of this specification, valid Message-ID values are those that have the format:

“<” id-left “@” id-right “>”

The *id-left* and *id-right* sections above MUST consist of 1 or more printable ASCII characters in accordance with RFC5322.

A valid Message-ID value MUST NOT contain any spaces. The spaces in the format line above are only used to illustrate the different sections of a Message-ID.

RFC5322 contains more detailed requirements for the format of Message-ID values. However, for the sake of simplicity and consistency, any value which matches the above requirements is considered valid for the purposes of this specification.

Example

Message-ID header line from email:	Message-ID: <CALckR-a8UDkRjO4xJyjd_s0GPxQWw@mail.gmail.com>
Value passed to MIH generator:	<CALckR-a8UDkRjO4xJyjd_s0GPxQWw@mail.gmail.com>
Generated MIH:	1de319c276884bd0c9e2f1621ada26cc

ADVISORY NOTES

The EDRM Duplicate Email Identification Specification is chiefly intended to facilitate cross-platform duplicate identification of RFC 5322 compliant email messages exchanged in eDiscovery and Disclosure irrespective of the form or forms of production; however, the Specification may also prove useful as a lightweight means of deduplication where limitations attendant to its use are well understood and accepted.

RFC 5322 STANDARD

As early as 1982, the standards for the format of "Internet Text Messaging" (better known as e-mail) required that each message transiting the Internet must contain a globally unique identifier. See, ***e.g.,*** RFC 822 https://datatracker.ietf.org/doc/html/rfc822, August 13, 1982. The latest revision of the standard, RFC 5322, reads:

> *The "Message-ID:" field provides a unique message identifier that refers to a particular version of a particular message. The uniqueness of the message identifier is guaranteed by the host that generates it. This message identifier is intended to be machine readable and not necessarily meaningful to humans. A message identifier pertains to exactly one version of a particular message; subsequent revisions to the message each receive new message identifiers.*
>
> > *Note: There are many instances when messages are "changed", but those changes do not constitute a new instantiation of that message, and therefore the message would not get a new message identifier. For example, when messages are introduced into the transport system, they are often prepended with additional header fields such as trace fields ... and resent fields .... The addition of such header fields does not change the identity of the message and therefore the original "Message-ID:" field is retained. In all cases, it is the meaning that the sender of the message wishes to convey (i.e., whether this is the same message or a different message) that determines whether or not the "Message-ID:" field changes, not any particular syntactic difference that appears (or does not appear) in the message.*

RFC 5322 Internet Message Format, October 2008 https://datatracker.ietf.org/doc/html/rfc5322

Hash of the Message ID

The EDRM Duplicate Email Identification Specification Committee ("The Committee") elected to hash the Message-ID for several reasons, among them to ensure that the EDRM MIH would be of a fixed length and composition (always 32 hexadecimal characters) and be case independent. Further, hashing the Message-ID supplies a degree of anonymity in consideration of the fact that Message IDs customarily incorporate domain names, which some parties may wish to shield from disclosure in rare circumstances.

This Specification was drafted for use with the most common email systems used in the last 10+ years including Microsoft Outlook/Exchange/M365 systems, Lotus/IBM Notes/Domino, webmail applications (*e.g.,* Gmail/Hotmail/Apple Mail) or any mail system where messages contain a globally unique, RFC-compliant Message ID.

CONSIDERATIONS WHEN MIH USED FOR DEDUPLICATION

Notwithstanding the requirement that Message IDs be "guaranteed" to be globally unique, the Committee identified scenarios where Message IDs were absent or were the "same" when the messages in which they presented were "different." In electing to deploy the EDRM MIH as a means of deduplication, the user must assess the potential for unexpected outcomes—whether manifesting as the failure to deduplicate functionally identical messages or the deduplication of items that are not, in fact, duplicates.

The scenarios listed below where the MIH on its own may not be adequate to perform deduplication SHOULD be considered. For example, for a particular matter, it might be decided to deduplicate using a combination of the MIH and the email Date (i.e. Sent Date & Time).

Some examples identified include, but are not limited to:

- **Draft messages without Message IDs**: Because Message IDs are generated by the email client program sending the message or by the first transmitting mail server, unsent "Draft" messages lack Message IDs. Accordingly, the Committee recommends that messages lacking Message IDs be deduplicated other than by use of an EDRM MIH. *Implementations should be designed so as not to produce an EDRM MIH when the Message-ID field is not validly populated.*
- **SPAM and Fraudulent Messages:** SPAM messages may be constructed to re-use Message IDs across different solicitations. Likewise, messages may be spoofed or forged resulting in the replication of the Message-ID of a valid message being repeated in a different or altered message. The decision to employ the EDRM MIH for deduplication should factor in this potential.

- **System Generated Emails:** The potential exists for non-compliant applications to generate emails with matching Message IDs (an example encountered in testing was a 3rd party Human Resources-type system that generates reminders with different dates and/or content with matching Message IDs). If non-compliant messages of this nature may be material to the investigation or matter, consider identifying duplicates utilizing field values such as Sent Date in addition to EDRM MIH.
- **Malformed or Corrupted Message IDs:** If the Message-ID is invalid because, *inter alia*, it's malformed, corrupted, or read incorrectly when hashed, the resulting EDRM MIH will be of no use except perhaps to identify messages with an identical erroneous value. This potential is not unique to the EDRM MIH, as it impacts all methods of deduplication.
- **Messages with Prepended or Appended Headers, Footers and Signatures:** Some email systems and servers prepend or append headers, footers and signatures to messages after the assignment of a Message ID. Thus, a message collected from a Sender's collection (i.e., from Sent Items) may not reflect the header, footer or signature assigned in transit (*e.g.,* by adding the word "[External]" to the Subject line). The consensus of the Committee was that, in most instances, reviewers would prefer to deduplicate all instances of the message despite the variation in header, footer or signature since the messages are "the same" in all material respects in most circumstances. In situations where the absence of a machine-generated header, footer or signature is deemed a material difference, alternate methods of deduplication should be employed.
- **Messages with Message Group and Alias Addressing:** Email systems typically support the sending of messages to groups of addresses using a group shorthand (e.g., "Board of Directors"). As well, email systems may use fully compliant email addresses *(i.e*, addressee@domain.com) or substitute the name of the addressee from the sender's contacts list. Because the Message-ID is the same for each, the differences in Group and Alias addressing may not be apparent when messages are deduplicated using the EDRM MIH.
- **Messages with BCCs:** Because a Message-ID is the same for a sent email including a blind copy (BCC) and each of the received emails, including the one received by the individual(s) blind copied, it is important to recognize that this may be material. The consensus of the Committee was that, in most instances, such as where 'daily updates' are BCC'ed to a group, reviewers would prefer to be able to identify all instances of the message and the BCC recipients, since the messages are "the same" in all other aspects. In situations where there is a BCC value, alternate methods of deduplication should be employed and users are cautioned to apply a secondary differentiating mechanism (*e.g.,* BCC) before deduplication. A similar application would apply where users wish to identify all recipients from a 'group email address'.

- **Messages with Stripped or Corrupted Attachments:** Because a Message-ID is a feature of an email message header, it will not change when a message is processed or "stripped" to remove attachments or when attachments have been altered or corrupted in processing. Accordingly, a stripped message or one with altered or corrupted attachments will nonetheless deduplicate against the originating transmittal when applying this specification. Where this is a concern, users are cautioned to apply a secondary differentiating mechanism (*e.g.,* file size comparison) before deduplication.
- **Messages with Time Anomalies:** Because operating systems and email client applications may resolve or display time information differently, for example, by using different time precisions, truncating values or rounding to whole minutes, messages presenting minor variations in time values but having the same Message IDs may be expected to deduplicate notwithstanding these time differences. The consensus of the Committee was that, in most instances, reviewers would prefer to deduplicate all instances of the message despite the modest variation in time values since the messages are "the same" in all material respects in most circumstances. In situations where the time variation is deemed a material difference, alternate methods of deduplication should be employed.
- **Items that are Not Email Messages:** This Specification has been drafted and tested solely for the purpose of deduplicating or identifying duplicate email messages. For other data present in mailboxes that may contain Message IDs (such as calendar or contact items), deduplication may or may not be feasible or reliable. This specification does not address that potential.

USE CASES

The Specification describes a means by which production sets supplying compliant Message-ID values can be deduplicated or duplicates identified, irrespective of the form or forms of production. For example, a static TIFF+ production of page images can be deduplicated or compared against robust native or near-native production formats like PSTs, MBOX, MSGs or EMLs. Historically cross-platform/cross-format deduplication or duplicate identification has been difficult or impossible because, when computing deduplication hash values, e-discovery service providers and tools tended to compare different components of messages in varying orders while normalizing the data in idiosyncratic ways. In addition, the deduplication hash values weren't routinely sought or produced in discovery.

The Committee anticipates that litigants would exchange EDRM Message Identification Hash values in load files as routinely as parties now exchange Bates numbers and file names. Such a modest evolution in good production practice will enable parties to save time and money by efficiently and inexpensively identifying duplicate messages obtained from other productions, parties and matters, even when the forms of production differ. **Further, by embracing the EDRM MIH, no service provider is obliged to change the way they deduplicate email messages internally.**

The primary use case allows parties receiving productions accompanied by EDRM MIH values to quickly determine if subsequent email productions contain duplicates of email previously seen, even when the forms of production have changed. It permits requesting parties to identify duplicates of email messages across production sets from the same party, across sets received from different producing parties and across different matters. Anyone who has ever tried to deduplicate a TIFF production set from one party against a PDF or native set from another understands the challenge of reliably doing so. This Specification makes it feasible as never before.

OTHER USE CASES:

- When opposing parties produce email messages that must be deduplicated or duplicates identified & grouped together against the client's own mail stores.
- Where jurisdictional restrictions and privacy regulations require client data to be hosted in two or more locations and parties need to identify and suppress duplicates without cross-border dissemination of private content or personally identifiable information.
- Where a custodian's email was previously collected and reviewed/coded in a collateral matter, even hosted by a different vendor, the EDRM MIH makes it possible to leverage those prior efforts and review only new and unique communications.
- The EDRM MIH supports identification of duplicates when foreign language encoding settings (for *e.g.,* Japanese) prompt different hash values when using conventional hash deduplication of text.
- Where a collection may span multiple email systems such as when a client may be partway through a migration between different email systems and may be running both systems in parallel such as a migration between Microsoft Exchange server on-premise and Microsoft Office 365, or between Exchange and Gmail etc, or even when a collection spans live mail servers and email archives.

SPECIFICATION NOTATION

Borrowing from Internet Standards RFC notation, this document occasionally uses terms that appear in capital letters. When the terms "MUST", "SHOULD", "RECOMMENDED", "MUST NOT", "SHOULD NOT", and "MAY" appear capitalized, they are being used to indicate particular requirements of this specification. A discussion of the meanings of these terms appears in RFC 2119.

EDRM EMAIL DUPLICATE IDENTIFICATION GUIDELINES (V1.0)

I. INTRODUCTION

The ***EDRM Message Identification Hash (MIH) Specification*** defines a process to identify duplicate emails across disparate formats and forms of data employed in electronic discovery and disclosure.

Currently no published solution exists for cross platform email duplicate identification. The EDRM Email Duplicate Identification Specification Committee ("Committee") crafted the EDRM Email Duplicate Identification Toolkit ("Toolkit") to facilitate cross platform identification of duplicate email messages.

The Committee anticipates that the use of EDRM MIH will lead to significant time and cost benefits, including:

- Smaller review populations, shorter review times and smaller hosting fees;
- Greater consistency in coding, leading to improved predictive coding scoring and less need to reconcile disparate coding for duplicates;
- Greater insight into other parties' productions with less time and effort;
- The ability to repurpose and leverage previous work product for future matters;
- Increased flexibility in allowing data to remain in multiple databases and locations while data minimization techniques are applied before consolidating the data into a central location; and
- Greater ease in moving matters across platforms or service providers.

II. TOOLKIT

The EDRM Email Duplicate Identification Toolkit comprises:

a. **The Specification**
A succinct, technical specification with advisory notes geared to software developers. The latest version of the **Specification** should be read in conjunction with the latest version of these ***Guidelines.***

b. **Guidelines**
This document is the EDRM Email Duplicate Identification Guidelines which outlines the objectives, methodology, potential use cases, advantages, and usage considerations of the Specification. These Guidelines are intended for use by those who implement the Specification, including parties and counsel, vendors and service providers and regulators and courts. ***Parties and Counsel*** are encouraged to produce and provide EDRM MIH values to facilitate efficient cross platform identification of

duplicate messages. ***Vendors and Service Providers*** are encouraged to support the calculation of EDRM MIH values in their production workflows, without the need to change how they deduplicate email messages internally. ***Regulators & Courts*** are encouraged to include the EDRM MIH as a datapoint defined in exchange protocols and practice notes.

c. **Whitepaper**
A practical, non-technical introduction to use of the EDRM MIH.

d. **Tools**
Data and utilities to support use of the EDRM MIH Specification:

1. **Test Data Set –** A corpus of emails enabling testing and verification of EDRM MIH implementations.
2. **Small Dataset MIH Calculator–** Excel-based tool to generate EDRM MIH values for small sets of Message-IDs.
3. **Open-source code to extract and calculate EDRM MIH values from Outlook PST containers.**

III. OBJECTIVES

During discovery or disclosure, it is often useful to identify duplicate emails in data exchanged between parties. Duplicate identification delivers many benefits including enabling legal teams to distinguish emails already reviewed from newly seen or -produced messages. Using prevailing approaches, it is difficult or impossible to identify duplicate emails across data sets produced by different parties using different tools and/or in different formats without reprocessing native data.

The difficulty arises because, although vendors use *similar* methods to detect email duplicates, there are *nuanced differences* between their methods. For example, there may be differences in the choice and order of metadata fields to be hashed, date and time formats, the treatment of blank spaces, the incorporation of field separator symbols and the sequence in which the fields are input to the hashing algorithm. Such variations produce different hash values for the same email. Therefore, while current approaches effectively identify duplicates within datasets processed entirely by each vendor's own platform, they do not enable duplicate identification across data processed by multiple vendor platforms. The EDRM MIH Specification addresses this challenge by facilitating cross platform identification of duplicate emails.

IV. METHODOLOGY

The EDRM MIH Specification takes a simple but effective approach to cross platform email duplicate identification by computing the MD5 hash value of a standard, obligatory email metadata field called the *Message-ID*.

Cross-platform/cross-format deduplication or duplicate identification has been difficult or impossible because, as noted above, eDiscovery service providers and tools tend to use different methods when computing deduplication hash values. Moreover, deduplication hash values weren't routinely exchanged in discovery or provided as part of disclosure. By computing and producing the hash of compliant Message-ID values, duplicate messages may be identified irrespective of the form or forms of production. For example, a static TIFF+ production of page images can be deduplicated against robust native or near-native production formats like PSTs, MBOX, MSGs or EMLs.

The Committee anticipates that litigants would exchange EDRM MIH values in load files as routinely as parties now exchange Bates numbers and file names. It is also anticipated that parties would provide EDRM MIH values in load files when producing documents to third parties, such as regulators, in the context of investigations and document requests. Such a modest evolution in good production practice will enable parties to save time and money by efficiently and inexpensively identifying duplicate messages obtained from other productions, parties and matters, even when the forms of production differ.

Cross platform deduplicate identification using the MIH may be used on data sets that have already been deduplicated by proprietary deduplication processes.

V. USE CASES

The primary use case allows parties receiving productions accompanied by EDRM MIH values to quickly determine if subsequent email productions contain duplicates of email previously seen, even when the forms of production have changed. It enables requesting parties to identify duplicates of email messages across production sets from the same party, across sets received from different producing parties and across different matters. Anyone who has ever tried to deduplicate a TIFF production set from one party against a PDF or native set from another understands the challenge of reliably doing so. This Specification makes it feasible.

The Committee identified the following non-exhaustive list of additional use cases:

- Where jurisdictional restrictions and privacy regulations require client data to be hosted in two or more locations and parties need to identify and suppress duplicates without cross-border dissemination of private content or personally identifiable information.
- Where a custodian's email was previously collected and reviewed/coded in a collateral matter, even hosted by a different vendor, the EDRM MIH makes it possible to leverage those prior efforts and review only new and unique communications.
- Where foreign language encoding settings (for ***e.g.,*** Japanese) prompt different hash values when using conventional hash deduplication of text.

- Where a collection may span multiple email systems, such as when a client may be partway through a migration between different email systems and running both systems in parallel (such as a migration between Microsoft Exchange server onsite and Microsoft Office 365 in the Cloud, or between Exchange and Gmail etc.), or even where a collection spans live mail servers and email archives.
- Where a collection contains emails with headers, footers and signature text blocks have been automatically added in transit and such emails need to be treated as duplicates.
- Where a collection contains emails where an alias name (eg. "Board of Directors") is substituted for full recipient email addresses and such emails need to be treated as duplicates.
- Where a collection contains substantively identical emails that have minor variations in time (because email applications may resolve time values differently).

VI. CONSIDERATIONS FOR USE

In electing to deploy the EDRM MIH as a means of deduplication, the user must assess the potential for unexpected outcomes—whether manifesting as the failure to deduplicate functionally identical messages or the deduplication of items that are not, in fact, duplicates. For example, the Committee has identified scenarios where Message-IDs were absent entirely from emails, or where the same Message-ID appeared in emails containing differences.

The scenarios listed below are those the Committee has identified as potentially unsuited to the use of the EDRM MIH as the sole means of duplicate identification. In these cases (and, potentially, others yet to be identified), alternate duplicate identification techniques may be more suitable; for example, by augmentation of the EDRM MIH approach with additional metadata fields, such as date and time sent or message size:

- **Draft messages without Message-IDs**: Because Message-IDs are generated by the email client program sending the message or by the first transmitting mail server, unsent "Draft" messages lack Message-IDs. Accordingly, the Committee recommends that messages lacking Message-IDs be deduplicated other than by use of an EDRM MIH. *Implementations should be designed so as not to produce an EDRM MIH when the Message-ID field is not validly populated.*

- **SPAM and Fraudulent Messages:** SPAM messages may be constructed to re-use Message-IDs across different solicitations. Likewise, messages may be spoofed or forged resulting in the replication of the Message-ID of a valid message being repeated in a different or altered message. The decision to employ the EDRM MIH for deduplication should factor in this potential for a spam email to have the same Message-ID for each recipient email This, may deliver benefits in cases where spam needs to be rapidly identified and/or discarded.

- **System Generated Emails:** The potential exists for non-compliant applications to generate emails with matching Message-IDs (an example encountered in testing was a third-party Human Resources-type system that generates reminders with different dates and/or content with matching Message-IDs). If non-compliant messages of this nature may be material to the investigation or matter, consider identifying duplicates using field values such as Sent Date in addition to EDRM MIH. The use of a common EDRM MIH for different system generated emails may deliver benefits in cases where such emails need to be rapidly identified and/or discarded.

- **Malformed or Corrupted Message-IDs:** If the Message-ID is invalid because, *inter alia*, it's malformed, corrupted, or read incorrectly when hashed, the resulting EDRM MIH will be of no use except perhaps to identify messages with an identical erroneous value. This potential is not unique to the EDRM MIH, as it impacts all methods of deduplication.

- **Messages with Prepended or Appended Headers, Footers and Signatures:** Some email systems and servers prepend or append headers, footers and signatures to messages after the assignment of a Message-ID. Thus, a message collected from a Sender's collection (i.e., from Sent Items) may not reflect the header, footer or signature assigned in transit (***e.g.,*** by adding the word "[External]" to the Subject line). The consensus of the Committee was that, in most instances, reviewers would prefer to deduplicate all instances of the message despite the variation in header, footer or signature since the messages are "the same" in all material respects in most circumstances. In situations where the presence or absence of a machine-generated header, footer or signature is deemed a material difference, alternate methods of deduplication should be employed.

- **Messages with Message Group and Alias Addressing:** Email systems typically support the sending of messages to groups of addresses using a group shorthand (e.g., "Board of Directors"). As well, email systems may use fully compliant email addresses (***i.e***, addressee@domain.com) or substitute the name of the addressee from the sender's contacts list. Because the Message-ID is the same for each, the differences in Group and Alias addressing may not be apparent when messages are deduplicated using the EDRM MIH.

- **Messages with BCCs:** Because a Message-ID is the same for a sent email including a blind copy (BCC) and each of the received emails, including the one received by the individual(s) blind copied, it is important to recognize that this may be material. The consensus of the Committee was that, in most instances, such as where 'daily updates' are BCC'd to a group, reviewers would prefer to be able to identify all instances of the message and the BCC recipients, since the messages are "the same" in all other aspects. In situations where there is a BCC value, alternate methods of deduplication should be employed and users are cautioned to apply a secondary differentiating mechanism (***e.g.,*** BCC) before deduplication. A similar application would apply where users wish to identify all recipients from a 'group email address'.

- **Messages with Stripped or Corrupted Attachments:** Because a Message-ID is a feature of an email message header, it will not change when a message is processed or "stripped" to remove attachments or when attachments have been altered or corrupted in processing. Accordingly, a stripped message or one with altered or corrupted attachments will nonetheless deduplicate against the originating transmittal when applying this specification. Where this is a concern, users are cautioned to apply a secondary differentiating mechanism (***e.g.,*** file size comparison) before deduplication.

- **Messages with Time Anomalies:** Because operating systems and email client applications may resolve or display time information differently, for example, by using different time precisions, truncating values or rounding to whole minutes, messages presenting minor variations in time values but having the same Message-IDs may be expected to deduplicate notwithstanding these time differences. The consensus of the Committee was that, in most instances, reviewers would prefer to deduplicate all instances of the message despite the modest variation in time values since the messages are "the same" in all material respects in most circumstances. In situations where the time variation is deemed a material difference, alternate methods of deduplication should be employed.

- **Items that are Not Email Messages:** This Specification has been drafted and tested solely for the purpose of deduplicating or identifying duplicate email messages. For other data present in mailboxes that may contain Message-IDs (such as calendar or contact items), deduplication may or may not be feasible or reliable. This specification does not address that potential.

VII. FREQUENTLY ASKED QUESTIONS

What is an email Message-ID?
It is a globally unique identifier allocated to every RFC 5322- compliant email message. As noted above, the Committee has identified certain scenarios where Message-IDs were absent from emails, or where the same Message-ID appears in emails with certain differences (eg. emails with prepended or appended headers, footers or signatures, emails with message group and alias addressing, and emails with time anomalies).

Why use the EDRM MIH alone, without any other metadata fields?
Prevailing approaches use multiple email metadata fields to identify duplicates, with each vendor approaching the problem differently. This leads to cross-platform inconsistency.
By comparison, a Message-ID is like a 'lowest common denominator', a single, consistent, universally unique metadata field accessible to all software platforms.

Why hash the Message-ID?
Hashing converts the Message-ID value to a fixed length and composition (32 hexadecimal characters using MD5) and makes the EDRM MIH case-independent while supplying a degree of anonymity, (*eg.,* shielding domain names).

Will it work for items in mailboxes that are not emails?
No. The EDRM MIH specification is only suited to email messages. It is not intended for use with other mailbox data containing Message-IDs, such as calendar or contact items.

Is the Email Message-ID approach employed by the EDRM MIH superior to existing approaches to identification of duplicate emails?

Because each technology vendor approaches duplicate identification in different ways, existing approaches do not support cross-platform duplicate identification. The EDRM MIH method will often identify more duplicates more effectively than traditional methods because it tolerates minor, often insignificant, variations that can occur in multiple instances of the same email. Nonetheless, by embracing the EDRM MIH, no service provider is obliged to change the way they deduplicate email messages internally. The EDRM MIH is most likely to be used post-processing, that is, after a party has suppressed duplicates using internal processing methodologies.

Why should parties specify production of the EDRM MIH?
Routine production of the EDRM MIH will allow parties to identify more duplicates and will result in fewer duplicate emails in data from diverse sources, platforms and systems, and from multiple parties. This means everyone in the eDiscovery ecosystem benefits:

- Clients will save time and money;
- Courts and Regulators will identify key documents faster and with less effort;
- Law Firms and Service Providers will expend less time and cost on duplicative data. Further, the guidelines will provide greater flexibility in the choice of eDiscovery platform at any given stage of a case lifecycle; and
- Access to justice benefits as it will be less costly for parties to pursue and defend cases because a reduction in the number of duplicates reduces the cost and burden of discovery.

Is this a mandatory standard?
The EDRM MIH Specification is offered as a voluntary measure. If the decision is made to use the EDRM MIH, then certain steps are required to claim compliance. We anticipate the EDRM MIH will see rapid industry adoption because its use saves time and money while improving efficiency. Over time and in the face of broad adoption, certain jurisdictions may choose to incorporate these guidelines into their eDiscovery protocols and rules of procedure.