

**Technology Brief:**

**Clearwell Discussion Thread  
Analysis**



## Overview

Clearwell’s patent-pending algorithms dynamically link together all related messages into chronological threads which capture the entire discussion, including all replies, carbon copies, blind carbon copies, and forwards. By walking the discussion thread, you can quickly identify all the participants, and determine who knew what, and when. Clearwell’s algorithms not only analyze the email metadata, but also the contents of email messages and attachments. In addition to a graphical visualization of the thread shown in figure 1, Clearwell provides analysis of the thread’s key terms, conversation pairs, contributors, and attachments.

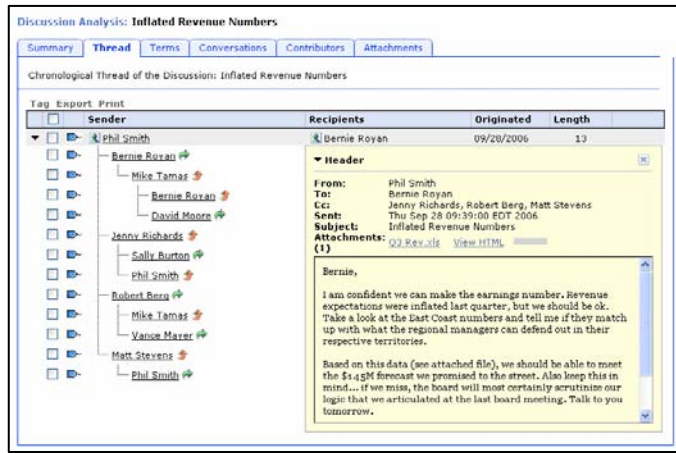


Figure 1: Screenshot of Discussion Thread shown in the Clearwell Email Intelligence Platform

This technology brief provides additional background on Clearwell’s Discussion Thread Analysis touching on three of the key processing elements shown in figure 2: Derived Email Processing, Duplicate Elimination, and Thread Construction.

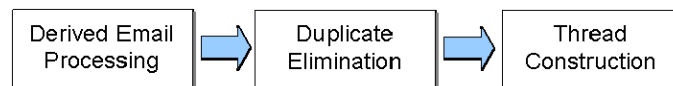


Figure 2: Key processing elements of Clearwell Discussion Thread analysis

## Derived Email Processing

One of the distinctive capabilities of the Clearwell Email Intelligence Platform is Derived Email Processing (DEP). DEP performs deep analysis within emails and extracts information that might otherwise be missed in a more surface-level analysis. DEP even allows Clearwell to identify messages that may have never formally resided in mail sources that are being analyzed. Such analysis is a critical step in accurately identifying all emails belonging to the same thread.

When a sender forwards or replies to an email, the original message text is automatically included within the body of the new email. As shown in figure 3, Microsoft Outlook® uses “-----Original Message-----” as the default prefix to start the

included message. It is also possible to configure each line of included text to start with a delimiter such as “>”.

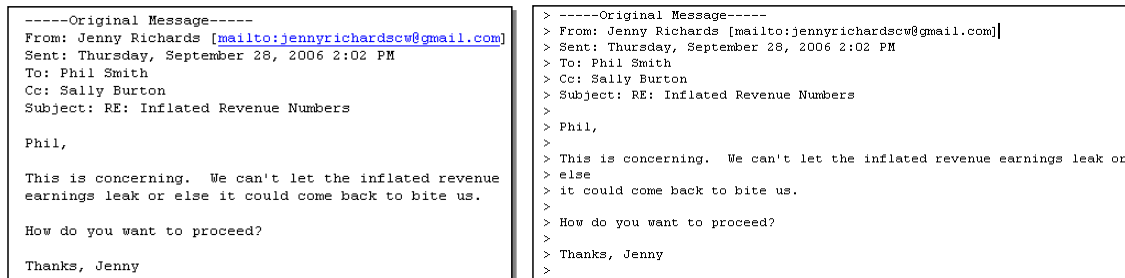


Figure 3: Examples of default prefixes for included email messages in Microsoft Outlook

DEP parses the body of every email, extracts included messages, and classifies these messages as Derived Emails. DEP’s advanced recognition features allow it to account for changes in message prefixes such as “-----Original Message-----” or “>” and account for their notation as forward/reply levels change. For example, when using a prefix such as “>”, it is common to add an additional prefix to notate a change in forward/reply level. The second forward/reply would use a prefix of “>>”, the third would use “>>>”, etc.

Figure 4 illustrates an example of how DEP would derive messages within an email. In this example, two derived messages are found within the email.

Building upon this example, Figure 5 illustrates how performing only a surface level analysis would yield significantly different results when compared to using DEP. If only an email’s header information is examined, the only message found would be the email from Sally to Mark. DEP reveals the series of events leading up to Sally’s forward to Mark.

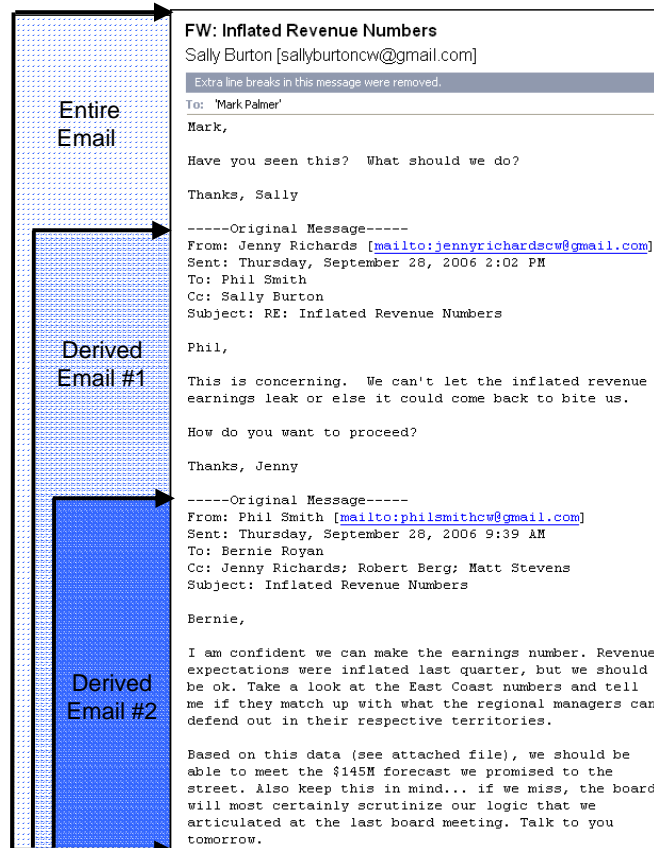


Figure 4: Derived email detection in the Clearwell Email Intelligence Platform

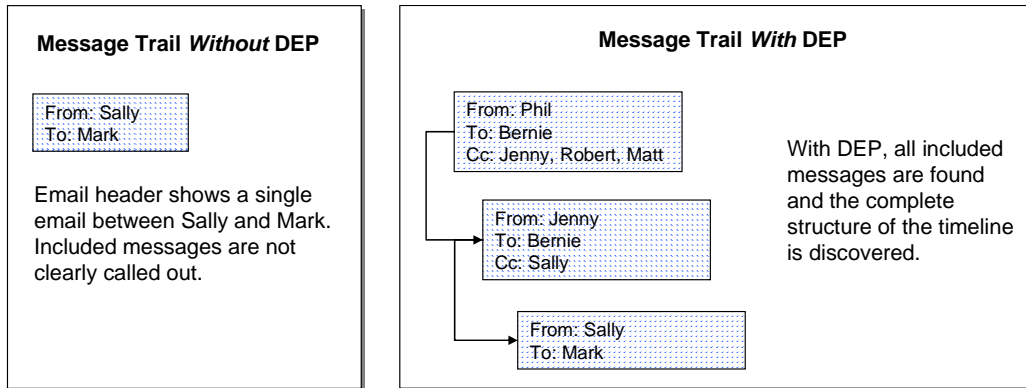


Figure 5: DEP accurately detects each email in the discussion

## Duplicate Elimination

Once Derived Email Processing has discovered all included messages, Clearwell's Duplicate Elimination is performed across the email corpus. Duplicate Elimination identifies duplicate emails and treats them as a single entity, eliminating redundant review of the same content by multiple people.

Traditional de-duplication techniques often apply a hashing algorithm (such as MD5) to generate a "strict" hash across an entire email. Unfortunately, such techniques can result in a large number of duplicate emails being unnecessarily included in the result set because, as the same email traverses across multiple email servers and mailboxes, minor formatting changes are often applied to the email. For example, spaces are often replaced for tab characters, and emails are often reformatted between plain text and html. Implementing a "strict" hash of the email will miss these differences preventing the same emails from being identified as duplicates.

Clearwell's Duplicate Elimination employs an advanced hashing algorithm that considers all of the unique properties of an email when making duplicate elimination decisions. In addition to taking into the account significant properties of the header, body, and size of the email, Duplication Elimination analyzes properties such as:

- Formatting differences between HTML, Rich Text, and Plain Text
- Certain content changes such as replacing a space for a tab character
- Addition of disclaimer text
- Conversion of attachments into attachment-file-names
- Differences in timestamp of emails for derived emails (e.g., forwarded/replied email don't include seconds in the timestamp)

This results in more effective duplicate elimination, decreased review time, and more accurate Discussion Thread Analysis.

## Thread Construction

Once Derived Email Processing and Duplicate Elimination have been performed, Thread Construction builds all final threads by analyzing relationships between emails.

Thread Construction performs the following steps:

1. **Subject Header Examination:** Thread processing begins with a subject header examination. Email subject fields are “cleaned” to remove forwarding and replying characters such as “FW” and “RE”. Subjects are then examined for similarity and potential thread candidates are grouped together into candidate pools.
2. **Derived Email Identification:** As described in detail in the section on Derived Email Processing, Derived Emails are messages found as included text within the body of an email. An email containing one or more Derived Emails is often part of a thread with the Derived Emails it contains. Emails within each candidate pool containing Derived Emails are identified and partial threads are established.
3. **Content Search:** Within each candidate group, email content searches are performed to further narrow down the candidate pool. Emails with matching content will remain in the candidate pool. These emails are compared against threads established in step 2 and consolidated if necessary.
4. **Chronological Examination:** Once all candidates for a thread have been identified, a chronological examination is performed. Email timestamps are normalized across time zones and logic is applied to account for timestamps differences that can occur due to machine time variations. The originating email is identified and the chronological order of the rest of the candidates is established.
5. **Visualization Processing:** Once the thread has been established, additional analysis takes place. Graphical visualizations of the thread are built, key terms are extracted, conversation pairs and key contributors are identified, and attachment analysis is performed.

## Summary

Clearwell's Discussion Thread Analysis consists of three key processing stages to detect, construct, and present accurate discussion threads. Clearwell's Derived Email Processing first detects all email within forwards, replies, etc.—even finding messages that were previously invisible to the email repositories being analyzed. Second, Clearwell's advanced Duplicate Elimination techniques consider numerous properties of email to deliver a much higher de-duplication rate than other solutions. The final stage, Clearwell's Thread Construction, uses a multi-step approach that detects even the most subtle relationships between emails in order to ultimately build and display each discussion thread. As a result of this patent-pending process, Clearwell's Discussion Thread Analysis assembles accurate discussion threads, thereby enabling a more efficient and confident email review and analysis process.