

# Clearwell Systems

## *Technical Note: Email & File De-Duplication*



January, 2007

---

### **Overview**

Clearwell employs de-duplication algorithms to identify duplicate emails, attachments, and loose files. The de-duplication algorithms are based on computing a message digest hash for each document (email, attachment, or loose file) which is then stored in the Clearwell master index. The message digest currently used is the MD5 algorithm.

To identify duplicates, hashes for new documents are computed and compared with the hashes already stored in the master index. In order to accommodate legitimate variations in documents between two originals, the content for computing the hash is carefully selected and outlined in the sections below.

### **Email**

The email hash is computed using the following email properties:

- Sender's email address
- To list email addresses in sorted order
- Cc list email addresses in sorted order
- Bcc list email addresses in sorted order
- Alpha-numeric characters from the subject of the email (removing spaces, tabs etc.)
- Time the email was sent after converting the time to UTC
- Body content

### **Attachments and Loose Files**

For attachments and loose files, an MD5 hash value is computed against the full content of the document. Unique document content is indexed only once. If two documents are equal in content but have different meta-data (such as attachment filename), the content is indexed once for search purposes and its meta-data properties are tracked independently. Meta-data properties for all attachments and loose files are made available for searching and are independent of the content de-duplication.