

# Data Deduplication Explained

## A GEON Legal Solutions White Paper

---

20<sup>th</sup> Aug 2010



“Providing the best litigation support service to you and your firm”

## **Table of Contents**

Introduction _____	<b>3</b>
What is Deduplication? _____	<b>5</b>
What is Near-Duplication? _____	<b>6</b>
AD Summations Enterprise Deduplication Tool _____	<b>7</b>
<i>How does it work?</i>	
<i>Specific Terminology</i>	
How can Deduplication benefit Litigation? _____	<b>9</b>
What other Advantages has Data Deduplication for my Firm? ____	<b>11</b>
Summary _____	<b>12</b>
Bibliography _____	<b>13</b>

## **Introduction**

There was no global recession in data growth in 2009, according to IDC research the worldwide volume of digital data grew by 62% between 2008 and 2009, to nearly 800 billion gigabytes. Considering that the world sends over 60 billion e-mails daily, that 90% of all documents generated today are electronic and that a single hard drive can store the equivalent of 40 million pages it is not surprising that the volumes of Electronically Stored Information (ESI) continues to expand at this alarming rate.<sup>1</sup> Therefore, litigation costs in relation to preserving, collecting, reviewing and producing this electronic data are understandably growing daily.

This enormous amount of electronic data has created significant obstacles for solicitors and their clients. Moreover, ESI can be altered, corrupted or lost and may be duplicated and dispersed with the click of a button. Today legal professionals must not only understand substantive and procedural law, they must also understand legal technology and data preservation retention policies in order to navigate through the growing world of ESI and to gain a competitive advantage in the industry.<sup>2</sup>

Amendments to Rule 12 (1) of Order 31 of the Rules of the Superior Courts, which have traditionally governed all procedures in the High Court in relation to the discovery process in the Irish Courts, have further complicated the e-discovery process. These rules have outlined that where an order for discovery includes ESI the court can order that the information is provided in searchable format or may order that an independent expert carry out the inspection and search for the relevant information. Applications to the court by way of Notice of Motion shall now specify the precise categories of documents in respect of the discovery sought.

As well as these strict regulations the nature of ESI itself poses unique issues for litigants. Unlike paper documents stored in filing cabinets, banker's boxes and warehouses, ESI resides in Word and Excel files, e-mails, instant message conversations, CDs, DVDs flash drives, thumb drives, voice mails, digital photographs, CAD files and other media, making data hard to track down. <sup>1</sup>

There is a growing gap between the amount of digital data being created and the amount of available storage. The carrying costs associated with storing and managing all that ever-increasing amount of data on disk or tape can be cut dramatically by deduplication.<sup>3</sup> First-time users of our software tend to be very surprised by the sheer volumes of duplicate data discovered within a case.

## **What is Deduplication?**

The data deduplication market is currently worth an estimated \$1.2 billion<sup>4</sup> and in this current market downturn, where companies try to prevent purchase of more software, it comes as no surprise that many organisations, including legal firms, are increasingly looking at deduplication to manage their existing ESI.

Data Deduplication is the process of comparing electronic records based on their characteristics and eliminating redundant data, or as the name implies, removing duplicate files from a set of data, it is often called “intelligent compression” or “single-instance storage”. It works by eliminating redundant data and ensuring that only one unique instance of the data is stored. This helps to simplify the e-discovery process as it reduces the amount of electronic documents that need to be reviewed for counsel significantly.

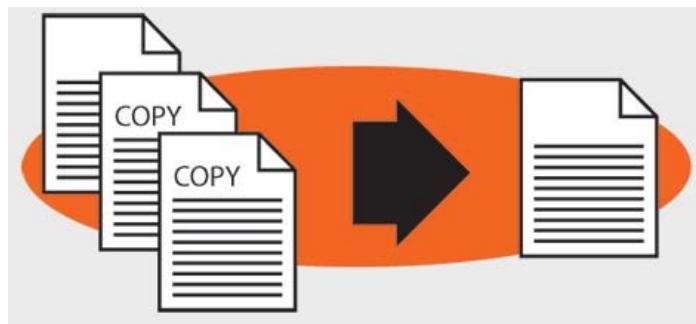
For example, if messages on an organization’s email server are being backed up and a mass email with an attachment was sent out to 100 employees of the organization, only one copy of the email is unique, since the other 99 instances are duplicates. With data deduplication solicitors will only have to review 1 email instead of 100 of the same document. However, all 100 emails can be restored if necessary. Additionally, if the email and attachment are 5MB, the total space used by the 100 emails is  $100 \times 5\text{MB} = 500\text{MB}$ . With data deduplication, the data backup will only take up 5MB.

The review process will be significantly less time consuming, also lowering risks where inconsistent treatment of documents may occur during human review.

## What is Near-Duplication?

The volume of duplicate documents in litigation discovery is overshadowed by the number of “near-duplicates” – significantly similar versions that differ by, for example, a few words or paragraphs<sup>5</sup>- this may be seen when comparing a draft document with its finalised version.

Unlike duplicate documents, it is often not adequate to simply remove all near-duplicates from the case as if this is done solicitors may lose the value of some very important information. If near-duplicates are grouped together it is possible to manage the review process such that the same reviewer is always reviewing similar documents. This helps to enable a more efficient and consistent review process while also ensuring consistent coding of similar documents. Moreover, the user is sure not to miss minor, yet potentially critical, differences that could possibly help the case.<sup>6</sup>



## **AD Summations Enterprise Deduplication Tool**

### **How does it work?**

AD Summation Enterprise allows users to export data from the database (.csv files) to a third-party tool, the Equivio analyzer. After an analysis takes place, the data can be imported to the database using the *CT Summation Data Manager (CTSDM)*.

The fields used to identify duplicate documents are populated during the import process. The information in related duplicate fields is used to speed up the document review process. Reviewers are able to review more documents in less time.

The document compare tool (*DOE Compare*) allows users to view all of the differences between a selected document and a Pivot within the same document set. In addition, users are able to apply bulk coding to all/or to a range of documents within the set, making the coding process more efficient and more accurate.

### **Specific Terminology:**

<b><u>Term</u></b>	<b><u>Description</u></b>
<b>Identical</b>	Two documents are identical if they are exactly the same.
<b>Exact Duplicate</b>	Two documents are exact duplicates if the textual content of the two documents is the same. A Microsoft Word file and the PDF version of that file are duplicates.
<b>Near-Duplicate</b>	Two documents are near-duplicates if there are "small" textual differences between the documents.
<b>EquiSort (GroupSort)</b>	EquiSort stores global information and groups documents into EquiSets (and e-mails into EmailSets).

**EquiSet  
(DocumentSet)**

EquiSets are groups of Exact Duplicates and/or Near-Duplicates. A document may belong to no more than 1 EquiSet.

*Note: If any two documents are assigned to one EquiSet, they will remain together in the same EquiSet, even if additional documents are added to the set, or if sets are merged.*

**Pivot Flag  
(PivotDocument)**

The Pivot Flag in each EquiSet is typically used as the baseline file for review. The Pivot document criteria are configurable options in the content analysis tool and are based on the criteria that best suits user needs. This is also referred to as the Prototypical Document.

*Note: The Pivot document can change when new documents are added to a set, once sets are merged.*

**Similarity  
(PercentSimilar)**

The percentage resemblance in comparison to the Pivot document.

Put simply, identical, duplicate and near-duplicate documents are determined by the AD Summation Enterprise deduplication software. Duplicates and near-duplicate documents are grouped together into an EquiSet and a Pivot Flag is selected in each EquiSet. A Pivot Flag is one document within an EquiSet that all duplicates or near-duplicates are compared against within certain criteria (which is variable depending on user needs) to measure their similarity in a percentage form.

All errors found during the export/import process of duplication analysis are displayed immediately and saved. Errors such as inconsistent treatment of responsive or privileged tags are displayed. Enabling you to eliminate redundancy in the document review process and significantly reducing the number of documents to be reviewed.

## **How can Deduplication benefit Litigation?**

The majority of commentators criticize Electronic Discovery for being a high cost within the litigation process however, since half the cost of a typical eDiscovery project is in document review; the potential here for cost savings is huge especially given that reviewing redundant email messages and near-duplicate documents is a huge component of that cost. By using data deduplication software solicitors have decreased the number of documents needed to be reviewed by as much as 90%, with 30-40% in the average case<sup>7</sup>, thereby lowering the risks, cost and time involved in production and review.

For solicitors conducting e-discovery, duplicated documents, near-duplicates and email threads are a well-known burden and a huge source of insufficiency in the litigation process, they add unnecessary risk and cost to the review and production of ESI. When employing large teams of solicitors and paralegals to conduct review without deduplication software, there is a great risk that individual reviewers will make decisions as to the privileged nature of documents that are not consistent with each other.<sup>5</sup>

Duplicate documents are marked appropriately and the Pivot Flag is determined so the reviewer can identify easily whether the rest of the duplicate or near-duplicate documents can be skipped or require a more detailed review depending on their percentage similarity to the Pivot Document, therefore irrelevant documents can be removed from the initial review process to ensure consistent and accurate document analysis.

Deduplication software allows counsel to receive data in a fraction of the time traditionally taken with human review. This process makes collected data available for pre-case assessment and analysis virtually as soon as it has been gathered. This helps to alleviate a lot of the burden placed on litigants as many cannot afford to review copies of every email and electronic file involved in e-discovery. Getting a better handle on the amount of data to be processed and reviewed, while reducing it significantly, allows counsel to be more prepared throughout the course of discovery.

Legal firms can eliminate compliance and e-discovery risks associated with using tapes – which are unwieldy to manage; can be lost, stolen or damaged; and may not be readily accessible. Deduplication makes it easier to meet government regulations and is compliant with the S.I No. 93 of 2009 in relation to the Irish Courts.

## **What other Advantages has Data Deduplication to my Firm?**

Traditional solutions for backing data are becoming increasingly uneconomical because they back up everything- including duplicate data files. Deduplication shows storage savings of up to 95%, dramatically reducing the disk capacity needed to store backup. Firms can achieve up to a 50% increase in server consolidation through more efficient backup. Organizations can reduce bandwidth utilization by moving up to 98% less data than traditional methods.<sup>8</sup>

Data deduplication can help litigation firms satisfy their return on investment requirements by managing their data growth, increasing the efficiency of storage and backup, reducing the overall cost of storage, reducing operational costs – infrastructure costs which require space, power and cooling and additionally reducing administration costs.<sup>9</sup>

## Summary

GEON Legal Solutions together with AccessData Group LLC and Equivio can achieve the same if not better accuracy than human review, but in a fraction of the time and at a fraction of the cost helping your firm gain a competitive advantage in the industry.

### Benefits:

- **Reduces Cost and Effort:** Reducing review and handling costs by 30 - 50%<sup>4</sup>. Enabling you to streamline the data review effort and focus on only the relevant documents – not wasting time and money reviewing documents of little or no importance to the case.
- Deduplication **reduces data volumes** so substantially (anywhere up to 90%) that it removes the need for tape or offline tape storage – accomplishing faster backups and recoveries.
- Achieve the level of **litigation readiness** you need to respond to discovery, in a timely and risk free manner. Isolating and prioritising the most relevant documents from huge document sets as well as allowing virtual suppression of redundant data can help solicitors cut directly to the information they need. Speeding up the time to results, **reducing the risks** of missing a critical piece of information due to oversights of critical data.
- **Consistent Treatment:** Deduplication introduces dimensions of **structure** into e-discovery, helping solicitors cope with data sets easily and make more informed decisions in the early case assessment stage and on which case strategy to pursue. Groupings also help in the delegation of particular types of documents - ensuring similar documents are assigned to the same analyst and coded consistently.

## Bibliography

---

- <sup>1</sup> Kane, Sally “E-discovery Explosion, E-discovery Growth and Challenges”  
<http://legalcareers.about.com/od/careertrends/a/ediscovery.htm> [Accessed: 18th August 2010]
- <sup>2</sup> See Note 1
- <sup>3</sup> Acronis Inc, “How Deduplication Benefits Companies of All Sizes”,  
<http://www.scribd.com/doc/21166838/How-Deduplication-Benefits-Companies-of-All-Sizes>  
[Accessed: 19th August 2010]
- <sup>4</sup> Nallayam, Radhika 15/5/09 - Channel World, Data Deduplication: More in Less,  
<http://www.channelworld.in/node/120> [Accessed: 19th August 2010]
- <sup>5</sup> Redundant Data, Equivio, Business Management Magazine  
<http://www.busmanagement.com/article/Redundant-Data/> [Accessed: 20<sup>th</sup> August 2010]
- <sup>6</sup> “Equivio> Near Duplicates”, Equivio, <http://www.equivio.com/product.asp?ID=5> [Accessed: 19<sup>th</sup> August 2010]
- <sup>7</sup> Lange, Michele C.S. & Nimsger, Kristin M – Electronic Evidence and Discovery: What Every Lawyer Should Know Now (Second Edition) [Accessed: 17<sup>th</sup> August 2010]
- <sup>8</sup> “Backup with Data Deduplication”, EMC<sup>2</sup> <http://uk.emc.com/solutions/samples/backup-recovery-archiving/backup-data-deduplication.htm> [Accessed: 19th August 2010]
- <sup>9</sup> “The Business Value of Data Deduplication”, SNIA Data Management Forum  
[http://www.snia.org/forums/dmf/programs/data\\_protect\\_init/ddrsig/Dedupe\\_Business\\_Value\\_V5.pdf](http://www.snia.org/forums/dmf/programs/data_protect_init/ddrsig/Dedupe_Business_Value_V5.pdf)  
[Accessed: 19th August 2010]



96 Upper Drumcondra Road, Drumcondra, Dublin 9, Ireland

Email: [info@geonlegal.com](mailto:info@geonlegal.com)

Web: [www.geonlegal.com](http://www.geonlegal.com)

Phone: (01) 8572882