



De-duplication

Answering your questions

SEPTEMBER 29, 2006

COMMONWEALTH LEGAL



Definition

- Removal OR identification of exact duplicate files from a data collection



Why?

- To save REVIEW time
- NOT to save processing \$\$



Implications

- If identified records are flagged and NOT Tiffed, the dupes will be unavailable for TIFF review, redaction and production
- Reviewers save time because dupes can be excluded from searches



Considerations

- When to remove
- When to identify



When to remove

- Same data exists on server and laptop
for same custodian
 - *E.g. Exchange server and PST email*
 - *Backup copies of "My documents" folder*
- Same data exists on multiple backup tapes
- Duplicate shared folders contain same files



When to identify

- Same file exists for multiple custodians
- Same file exists in different shared folders



Methods

- De-duplication by custodian
 - Feasible, included with processing
- Project-wide
 - Can be challenging, not as useful



Changing Search Scope

- 50Gb collection indexed but not numbered
- Searches exclude 20Gb
- If you need to go back to the 30Gb, need to use hash codes to dedupe
- Very difficult at this stage



Changing search scope

- 50Gb indexed and numbered (processed) for native file review
- Searches exclude 20Gb
- If you need to go back to the 30Gb, easy to de-dupe
- But you do need to pay for the excluded 20Gb (\$15,000)