Integrating Digital Forensics Techniques into E-Discovery Processing and Review

Dr. Gavin W. Manes, CEO
Gavin W. Manes, Ph.D.

- CEO, Avansic
- Doctorate in Computer Science
- Scientific approach to e-discovery
- Published over fifty papers on e-discovery, forensics and security, presented at hundreds of events
- Expert witness
Agenda

• Overview of E-Discovery Workflow
  – Preservation, Collection, Processing, Review, Production

• Digital Forensics
  – Comparison to E-Discovery
  – Common techniques that apply to E-Discovery
Who are the Players?

• **Clients/Corporations**
  – Organize and know where data is
  – Continue doing business – carefully

• **Law Firm**
  – Direct discovery
  – Plan and strategize EARLY
  – Choose technology partners [vendor, in-house, client]
  – Review and redact

• **ESI and Digital Forensics vendor**
  – Communicate well with parties
  – Recommend best course of action based on experience
  – Manage expectations
  – Execute project plan as efficiently as possible
E-Discovery Workflow Pre-Review

• Collection
  – Interview client and if they have IT, make sure to talk with them
  – Preserve as much as possible and Collect from that preservation
  – Use Chain of Custody documentation

• Processing
  – You can choose to process only parts of the collected data
  – Dedupe, email thread, extract metadata and text, etc.

• Search/ECA
  – Have well-crafted search terms ready ahead of time
  – Run searches for “potentially privileged” at the beginning to remove them from first-pass review
E-Discovery Review and Production

• Coding
  – Use mass coding to your advantage in a first pass review to exclude junk email addresses, spam, whitepapers, etc.
  – Set up a coding panel that forces people to be specific
  – Don’t overload with coding tags
  – If you mark as privileged, must say why it makes privilege logs much easier

• Production and Exports
  – Either have vendor do electronic productions or you; not both
  – Follow the agreed upon format and keep a copy for yourself
  – Make sure you get metadata that allows for deduplication
    • Conversation index, description of the hash methodology
  – Get an updated custodian lists for deduplicated data for each production
E-Discovery & Digital Forensics

• **Shared goal**
  – Locate digital evidence to support investigations or fact discovery

• **E-Discovery**
  – Content-based
  – “What” of a scenario

• **Digital Forensics**
  – Context-based
  – “How” and “when” of a scenario

• **Example**
  – E-discovery would find a key email
  – Forensics would show how it arrived at the computer, how often it was opened, if it was sent to another location, etc.
E-Discovery & Digital Forensics Con’t

• E-discovery & digital forensics are both ways to filter ESI

• E-discovery
  – Filtering and reducing a document set based on search terms, duplication, custodian, dates, etc.
  – Also used to describe ESI Processing
  – Output is a set of files

• Digital forensics
  – Investigation into the active and inactive space of computers by an experienced examiner
  – Commonly associated with preservation
  – Output is a report plus a small set of relevant files
Forensic Capabilities

- **Computer investigations**
  - Email analysis
  - Internet history analysis
  - Software analysis
  - External device analysis (i.e., USB drives)
  - Registry MRU (most recently used) analysis
  - User activity timeline

- **Cell phone investigations**
  - Can potentially retrieve text messages, photos, call logs, data from apps, calendar information, email, internet browsing information
  - BUT – very dependent on the type of phone
Frequent Digital Forensics Tasks

- **Document Location and Authentication**
  - Key email produced by the opposing party is detrimental to the case
  - Search doesn’t find the email or anything similar
  - Investigator notices inaccurate header information (email dated after its supposed transmission)
  - Email might not be authentic/sent

- **User Activity**
  - Dates and times of file copies (especially to external media)
  - Deletion and wiping analysis with file recovery
  - Websites visited, USB drives used, files opened
  - Device location

- **All of these exams use metadata**
What is Metadata?

• “Embedded data” within an electronic document
  – Information may not be visible in a printout
  – Critical difference between paper and electronic documents

• Document information tracked by computer software

• Describes how, when, by whom a document was created, modified and transmitted (i.e. “a history of the document”)
  – Can include hundreds of other items
Sample Metadata Report

Forensics Techniques in E-Discovery
Types of Metadata

• Email Metadata
  – Hundred of elements: sender, cc/bcc, subject, date/time sent, received, attachments, forward history, etc.

• Document Metadata
  – Date and time created, modified, last accessed, document size, file location, etc.

• Application-Specific Metadata
  – Word processing documents track updates and edits; spreadsheets have calculations, etc.

• Operating System Metadata
  – System logs relating to files that are stored, last edited, user permissions, etc.
Metadata in Litigation and Internal Investigations

• **Additional source of evidence**
  – Information and history behind an electronic document

• **Useful for**
  – Authenticating documents
  – Resolving factual disputes
    • Who created a document, if email was sent or received, etc.

• **Valuable source of information for document review**
  – Automated filtering, coding, sorting, indexing, etc.
Metadata Considerations

• Electronic documents should be collected and preserved
  – As they are stored in normal course of business – including metadata

• Preserving does NOT necessarily mean producing in discovery which does NOT necessarily become evidence

• Some metadata may not be accurate
  – It is easily forged or accidentally changed
  – Must authenticate if relevant

• Consider whether metadata is crucial due to the costs and analysis involved
Carved, Deleted & Orphaned Data

• Data carving can locate fragments of deleted data
  – Intensive scanning

• This type of data is always considered in forensics investigations
  – They might have a place in the story about what happened
  – They might be the story

• Requires computer science expertise to interpret
Document Similarity

• Goal: Identify duplicates or near-dupes
• Of these three, which are most similar?
  – An apple a day keeps the doctor away
  – An orange a day keeps the doctor away
  – A doctor should never compare an apple to an orange
• What’s most important?
• Now extrapolate to hundreds of documents each with dozens of sentences
• Need an efficient way to find relevant info and numerical way to describe similarities
Document Processing for Analytics

• **How to prepare text for comparison**
  - Text normalization
    • Character encoding, remove non-alphanumeric characters
  - Tokenization
    • Split sentences into individual words or phrases
  - Removal of stopwords
    • Delete tokens/words/phrases that are irrelevant

• **Example text strings after processing and tokenization**
  - Apple, day, keeps, doctor, away
  - Orange, day, keeps, doctor, away
  - Doctor, should, never, compare, apple, orange
Document Comparison

• **Comparing tokenized documents**
  – Character based distance (Edit, N-Gram, etc)
  – Token based distance
    • Bag-of-words, bag-of-phrases, ordered multi-token methods
  – Latent semantic analysis or indexing

• **Example using our tokenized documents**
  – 1. apple, day, keeps, doctor, away
  – 2. orange, day, keeps, doctor, away
  – 3. doctor, should, never, compare, apple, orange

<table>
<thead>
<tr>
<th>Compare</th>
<th>Jaccard</th>
<th>2-Shingle</th>
<th>TFIDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 vs 2</td>
<td>0.67</td>
<td>0.6</td>
<td>0.75</td>
</tr>
<tr>
<td>1 vs 3</td>
<td>0.22</td>
<td>0.0</td>
<td>0.16</td>
</tr>
<tr>
<td>2 vs 3</td>
<td>0.22</td>
<td>0.0</td>
<td>0.16</td>
</tr>
</tbody>
</table>
Production and ESI Protocols

- Just having an ESI protocol and agreement is very important
- Include clearly defined goals in the protocol
  - i.e., “we want to collect (these kinds of) devices” and “(this kind) of data”
- Allow wiggle room to take advantage of changes in technology
  - New collection method (i.e., for social media)
- Don’t restrict parties from being able to view, share, or transfer the data to other parties and experts
- Ask someone with experience to look at it
Conclusion

- Both forensics and e-discovery are important parts of the toolkit
- Investigatory vs. process-based
  - Use each for their strength
- Goals can change mid-case
- Have both types of experts available
- Plan ahead for both production and review
  - Reasonable expectations for turnaround
- Project management is key
  - Save time and money
  - Eliminate duplicate review
- Use technology to your advantage
  - Sophisticated software exists
  - Used properly, it can ease the burden of e-discovery
AVANSIC
E-Discovery & Digital Forensics

avansic.com

Corporate Office
First Place Tower, Suite 1800
15 E. Fifth St, Tulsa, OK 74103