DEVELOPING AND MAINTAINING TRUSTWORTHY CONTENT

Mitch Cochran

Director of Information Technology

City of San Bernardino

May 17th 2017

Focus on the Content

- Maintaining the interpretation of the content when we digitize a physical document.
- How do we protect the authenticity and integrity of electronic documents?
- How do we protect the authenticity and integrity of content that is outside of a stand-along document file?

General Thoughts

- Document Mentality – laws are built around a hard copy
- Electronic System Considerations
  - What do IT Managers worry about?
  - Why does it matter to a Records Manager?
- Migration
  - New systems
  - New technologies for storage
  - New encryptions
- Documents vs Embedded Knowledge

Trustworthy Document

- As the document is entered, it is trusted. If anything changes, it is untrusted.
- Once you scan it you can’t go back and make it better later….
- It is difficult to add data or descriptions later.
  - Change Management Procedures – why, who authorized, etc.

California Secretary of State

22620.4. Official Document or Record Storage Using Electronic Technologies.

- To ensure that all electronic versions of official documents or records (including documents or records converted from hard copy or electronically originated documents or records) are stored and managed in a trusted system as required in Government Code section 12168.7(c), electronic content management systems implemented six months after the adoption of these regulations shall be designed in accordance with section 6.2 Recommended Project Steps and Activities of “AIIM ARP1-2009 Analysis, Selection, and Implementation of Electronic Document Management Systems,” approved June 5, 2009.

Focus on Content

- Our retention and management practices focus on the content
  - Laws are about the physical document

- Legal interpretation: a copy is allowed if the interpretation is not changed
- A fax is legal – but ugly…..

Managing Documents

- We have to look at the trade off between the best resolution and file size?
  - Files for general use
  - Files/images for magazine covers are higher resolutions
- Video
- OCR
- Color
- Image interpretation
- Enhancement

Trade-off between Resolution and Size

- Traditionally its about retention schedules
- From an IT perspective, we need to consider file management and in turn file size.
  - Disks are not infinite - cheap
  - More disks, more backups

- So what is the impact of higher resolution?

File Size

- So what is a page? 8.5x11? Legal?
- Consider DPI – scanned image
  - 200 x 200 = 40,000
  - 300 x 300 = 90,000 2x
  - 600 x 600 = 360,000 8x
- Consider page size –
  - Are there any legal docs out there?
  - D size – 2x3
  - E size – 3x4 – 1 to 2gb per page at 200dpi

What About Video?

- It is coming….
- Treat like a PDF
- Which file type as a standard?
- Large files
  - PD body cams – 2Gb per officer per shift for 2 years
  - Petabytes….
    - Kilo, Mega, Tera, Peta
  - Backups?
  - Just file transfer takes a while…..

Optical Character Recognition (OCR)

- Changing a scanned image into text
  - Create a bag of words
  - Highlight the misspelled like a Word document
  - Can decide to get 100% accuracy – not necessarily all files
  - Systematic errors
    - ll can be L L, H, one one, eye eye……
- Resolution matters
  - 200dpi – 80%
  - 300dpi – 90 to 96% - sweet spot – trade off of accuracy vs size
  - 400dpi – 93 to 98%
    - These numbers vary

Does Color Matter?

- Does color matter to the interpretation of the content?
  - Even red ink on an accounting report
- If you scan in grey scale, you may lose some interpretation
  - If colors are close, it could be hard to interpret.
- 8bit color – larger file size

Are There Images?

- What resolution will be needed to interpret and reproduce the images
  - In a perfect world, you know what you want to do with an image
  - DPI – do you need higher resolution?
    - Text is 200dp
    - PowerPoint on a screen is 96dpi
    - Magazine covers are 1400dpi
  - Compression – can you compress it?
- Once you compress, you can't go back. If you need to edit, you can't make it better later
  - Compression is a loss of information - Pixelation
  - Just like if you have a number: 4 you can't say it is 4.00000

So what about PDF’s?

- Small, direct from applications
- Consider – PDF / A = archival
  - All of the necessary fonts are included so it is stand alone
- Able to secure – password
  - I hope you remember it over time
    - My old pattern was the phone extension of the person that created it
    - Shh! – it’s a secret

Enhancing the Image - Tradeoffs

- The simple example – fix a fax by removing dots
- For documents –Kofax or software to remove dots, lines, etc.
- On video, as an example, some firms can remove glare so you can see inside the car windows.
- Picture the attorney asking you – was the document modified?
  - Instant doubt….
- Archivists say no modifications during scanning

Enhancing the Image – Secretary of State

- 5.4.2.5 Post-scanning processing Post processing may be used to provide image "clean-up" after the scanning and prior to indexing and final storage. This software generally performs de-speckling, de-skewing and other functions to improve the quality of the scanned image with limited operator intervention. Use of image “clean-up” and other post-scanning processing should only be used to improve legibility. Caution should be exercised when using these tools, as any material modification to the image may affect the ability to authenticate the document in a legal proceeding

Recommendations for Digitization

- 200 dpi – text content only
  - State Archivists specified 200dpi
- 300 dpi – if there are images
- 300 dpi with color (8bit)
  - Or grey scale

- Decision needs to be made by owner or by custodian
  - Probably done by the person scanning – need to give guidance..

General Procedural Issues

- Can we get rid of the paper documents?
- Cloud
- Meta-data

When can we get rid of the paper copy?

- Secretary of State –To ensure that every official electronic document or record is considered to be a true and accurate copy of the original information received and before the original copy may be destroyed, at least two (2) separate copies of the official document or record must be created on electronic media meeting all the conditions of a trusted system as identified in section 5.3.3 Trusted system and legal considerations of "AIIM ARP1-2009 Analysis, Selection, and Implementation of Electronic Document Management Systems," approved June 5, 2009, which is incorporated by reference in this section.
- So maybe we can get rid of paper???
  - Don't get excited….government is not that straight forward.

CA Secretary of State Trusted Document

- A trusted document management system ensures that all electronically stored information can be considered to be a true and accurate copy of the original information received regardless of the original format.
- The trusted document management system must ensure that at least two (2) separate copies of the electronically stored information are created meeting, at a minimum, all the following conditions:
  - (a) The trusted document management system must utilize both hardware and media storage methodologies to prevent unauthorized additions, modifications or deletions during the approved lifecycle of the stored information; and ASSOCIATION FOR INFORMATION AND IMAGE MANAGEMENT INTERNATIONAL 20 AIIM ARP1-2009 – Analysis, Selection, & Implementation of Electronic Document Management Systems (EDMS)

CA Secretary of State
Trusted Document

- (b) The trusted document management system must be verifiable through independent audit processes ensuring that there is no plausible way for electronically stored information to be modified, altered, or deleted during the approved information lifecycle; and
- (c) The trusted document management system must write at least one copy of the electronic document or record into electronic media that does not permit unauthorized additions, deletions, or changes to the original document and that is to be stored and maintained in a safe and separate location.

- Does (c) require a WORM drive?
- Would a backup count as a second copy?

When can we get rid of the paper copy?

- Vendors Interpretation:
  - Save to two originals – scan and save to two servers, not copy from server to server
    - Possible error could be introduced
    - Now have to manage two servers
  - If there are two different copies, which one is correct?

Cloud Issues – it is foggy

- Gee, everything can go to the cloud…..can't it?
  - Multiple copies in multiple places
    - Copy, that leaves some doubt
- Well, not according to current laws
  - Would need a WORM drive…paragraph (c)
  - No provider currently states that files can't change
- Does the document need to reside in the US?
  - Government clouds
- Access – Which cloud administrators can do what?

Meta-Data

- Meta Data is data about data
  - Who? How? What? When? Quality? How captured?
- Laserfiche calls the index – meta data
  - Not quite right, meta-data is more
- Once you have created the index
  - Difficult to add fields – You need to update all records
  - Delete fields – Normally you can’t delete index info
- Tradeoff – more fields for more customized searches vs the requirement to enter all of the data for all of the records
  - Extensive time taken to enter the data, quality controls
  - What will you truly need to search for?
    - Permits example – address or add owner, builder, engineer, etc.
    - OCR ? Zone OCR

Document Sharing

- As we get more advanced, we will have to worry about having consistent meta-data
- DOD:
  - Data, Information, and IT services will be considered trusted when they have provided sufficient pedigree and descriptive metadata for consumers to rely on them as an ADS, and comply with applicable information assurance and cyber security policies.

Electronic System Issues

- Data in Transit / Data at Rest
- Security on the server
- Audit
- Migration to new systems
- Ensuring a change

Data Custodian

- As the data custodian, you have a legal responsibility to protect the data.

- Due Diligence
  - Have you followed best practices to maintain the confidentiality and integrity of the information?
  - Hindsight by a judge or jury has 20/20 vision….

Data in Transit

- Data in Transit – Data that is in the communications

- We are use to talking about data being encrypted for communications. It is intuitive to keep communications hidden or encrypted.
  - HTTPS instead of HTTP
  - Most communications are HTTPS becoming by default

Data at Rest

- New focus is Data at Rest
- Data that is in the system
  - Databases, files
- Data needs to be encrypted while at rest.
  - Should someone copy a file, they can’t open it.
  - An application can open an encrypted stored document but the file by itself cannot be opened.

- Hackers get in but they can’t see anything….
- Laptops – should be encrypted so when they get lost….

Systems Storage – Security of Image File

- Document systems are both databases and image files
  - Typically on different servers
- The database system will return a file location, then the application gets the image
  - Only if the user has authorization within the application.
- The issue is that the application system does not manage the security of the server where the file is.
  - A user can look at the image storage server and directly touch the jpeg file bypassing the controls from the application program
  - You should manage security on the server to match what you have in the application

Audit

- How can you be sure that documents haven’t changed?
  - Intentional
  - Non-intentional
    - Save a corrupted document and you have 2 perfect copies of a corrupted document
- Date can be changed
- Size – a small change may not show up
- Hash Code

Hash Code

- It is a one-way mathematical function that generates a unique number based on the contents of the file.
  - It can’t be reversed engineered
  - Used to store passwords
- If the file changes, it will have a different hash code.
- Hash code is stored separately from document
  - Can’t be part of the document
    - If you calculate the hash, then add it to the document, it will be a new hash code since it will be the based on document + hash code
      - Which leads to a new hash code….
  - One more thing to manage

Systems - Migration

- Moving to a new media?
  - CD to DVD to local electronic to cloud
  - You will have to change media at some point
    - Do you have any 3 ½ disks left?, Zip drives?
- When a document is trusted, how do we assure it does not change when moving to a new media?
  - Like a backup – would need to run a verify
  - Check disk size
  - Prefer a hash code – focus on content of file

Migration to new Encryption

- So if the data is encrypted, how do we move it to a new encryption?
- AES – Advanced Encryption Standard is the current standard. Selected by the Federal government in 2001.
- As the next technology comes out, we need the capability to unencrypt the data, then encrypt with the new standard.
  - There is no new standard in the foreseeable future.

Systems

- Backups
  - We all believe in backups
- Are backups kept longer than the retention schedule?
  - So, you would have to retrieve it for a document request since you it is still there….
- How to separate out the documents by date, can you?
  - Probably not….

Retention Schedule Interpretation

- The document retention can be from the end of the document’s project
  - Grant ends in 2 years, the retention would be 2 years till it ends and then the 7 years on the schedule, so a total of 9 years from now.
- Goes back to the managing the files
  - When does the retention period end on which file?
  - Does that change how you manage it?

Content That Never Hits Paper

- PDF – even though it does not hit paper, it can be treated like it
- Summary Reports
  - Text search
  - Images
  - Maps
- Databases or knowledge embedded in applications

Applications / Databases

- What are you managing?
  - Reports – like document
  - Embedded data

- Some people keep old systems for retention
- How to ensure they are managed until the end of the retention period ends – IT will not know what the target date is.
- What happens if it is permanent data
  - Information Systems are not permanent

Application / Database Migration

- The City Clerk is not part of other departmental application migrations
  - But you are responsible for the output
  - A new finance system, a new fire/PD reporting system, a new permitting system
- Historical data is migrated to a new system
  - How do we ensure that content is maintained in its original form….
    - Bad news - You can’t

Application Migration

- Application migration decision, how much data to keep or migrate?
  - Higher cost for more data
  - 2 years, 3, 4, 7, all?

- This decision affects the document retention
  - If data is not migrated, the system must remain operational

Knowledge Migration

- The new system will have a migration plan for data or a “crosswalk”
  - Variables from one system are matched to the new system
    - Good to follow / easy to audit
  - Some data is dropped
  - Some data is changed – new field types, sizes
  - Some data is calculated
    - New calculations are different
- How do you want to document the changes for change management procedures?

Challenges

- Physical documents – digitization
- Electronic files
- Media changes
- System Migration
- Knowledge management

IT Managers Point of View

- Managers want a consistent way to handle documents regardless of their format
- Document clarity matters to the interpretation of the document
- Technology standards will evolve
- Authenticity – How can we establish that a document is authentic?
- Integrity – How to maintain a record over time
- Security – Encryption of the data at rest and securing direct access to the documents

IT Managers Point of View

- Physical Control – Are the documents located locally or in the cloud? Who can see them?
- Migration to new technologies – There need to be a capability to unencrypt and then move to a new encryption standard
- Document sharing among organizations
- Retention Period – what are the organization's retention policies and are they being followed.

Thanks

What are your thoughts?

What did I forget to mention?

Mitch Cochran

Cochran_mi@sbcity.org