

Can records Management be automated?

James Lappin
Thinking Records
blog: www.thinkingrecords.co.uk
jhlappin@gmail.com
@Jameslappin

ABSTRACT

This paper examines the reasons why sections of the recordkeeping world (and in particular the US National Archives and Records Administration) are looking to encourage automated approaches to records management. These approaches are aiming to reduce the burden of records management on end users.

Automated approaches are in sharp contrast to the efforts of most records management programmes over the last 15 years to give individual end users the responsibility for the key records management tasks of selecting and filing records

The paper outlines the different automated approaches on offer and concludes that whilst each of them has merit, none of them yet provides a fully scaleable solution to records management in organisations. In particular there is no automated solution currently available to tackle the problem of the build up of large scale e-mail aggregations in the form of e-mail accounts on servers and in e-mail archives.

As well as evaluating the automated approaches available from vendors in the content management space, this paper also looks at ways in which organisations can configure the way they manage e-mail with a view to having e-mails accumulate in manageable individual or team correspondence files rather than in unmanageable individual e-mail accounts.

The paper discusses ways three examples of records management systems that have worked well - two from the digital age and one from the paper age, and identifies two common denominators -

- the fact that they each involve some sort of intervention to control and filter the communication channels by which the business whose records they capture is conducted.
- the fact that they capture records that are referred to and relied upon by the people carrying out the work that the system records

Keywords

Records management

1. INTRODUCTION

Records are multi-faceted. They have many potential users.

We may think of a records continuum spreading out in each direction from the person(s) carrying out a piece of work, to their immediate colleagues and line management chain, and then, embracing people further removed in time and or/space - the successor(s) to the person(s) carrying out the work; auditors; legal and compliance colleagues. Depending on the nature of the work the continuum may stretch to external stakeholders such as customers, clients, regulators and citizens; and on further to

archivists and the future generation of researchers and historians that they serve.

The purpose of records management is to design systems that capture records in a requirements of all or most of the stakeholders on this continuum.

The last time records managers were able to design systems at a corporate scale that met the needs of all or most stakeholders was in the paper age, before the coming of e-mail.

The failure of organisations in the post e-mail age to design records systems that meet the needs of all stakeholders has meant that some stakeholders have taken to advocating systems that meet solely their own particular needs, whilst neglecting the needs of other stakeholders on the records continuum. For example:

- The need of an organisation's legal Counsel to respond more quickly to litigation requests might drive the implementation of e-discovery software and/or an e-mail archive.
- The need of the National Archives and Records Administration of America to capture for historians a record of the correspondence of senior federal officials drove them to rewrite their e-mail policy(1) and invite Federal Agencies to preserve and transfer the e-mail accounts of key staff.

These are examples of mono-faceted rather than multi-faceted approaches to recordkeeping:

- e-discovery systems have an incredibly powerful indexing and search capability. They enable a legal counsel to create a search string to pull back documents or e-mails created, sent or received by particular named individuals, within a particular time frame and which include particular words or phrases. But the organisation cannot allow anyone outside of their legal/compliance team to use that search facility. This is because it searches dark data - in particular data in e-mail accounts. To allow all colleagues to search such data would lead to unethical breaches of privacy and confidentiality.
- NARA's capture of significant e-mail accounts may help historians in 75 years' time, but it won't help the immediate colleagues and successors of those post holders. These e-mail accounts will be inaccessible to them unless the organisation can find a way to reliably filter out private and confidential correspondence

2. THE THREE AGES OF RECORDS MANAGEMENT

2.1 The changing position of the end-user in records management practice

In the days before e-mail the best records management systems were built on the belief that the capture of records was too important to be left to end-users. This belief held that it was important that officers/officials did not have the choice of which communications/documents arising from their work were and were not captured as a record.

One of the purposes of a records system is to hold those officers/officials to account for how they conduct their work (another is to enable those officers/officials to defend how they had conducted that work). If they can choose what goes onto the records then they can choose to leave off the record any communications/documents that could be detrimental to them.

These beliefs changed after the introduction of e-mail in the mid 1990s. Since the introduction of e-mail organisations have typically stated in their records management policy that it was the responsibility of each individual employee to capture and maintain a good record of their activities. This was often stated in moral terms - individuals were employed by the organisation to do a job and they therefore had a duty to leave behind a good record of that work

Since 2007 an information governance view has emerged that individual knowledge workers are too inconsistent and poorly motivated to perform records management tasks well, and that organisations would be better off finding ways to automate records management.

To find out the reasons for these sudden reversals in ideas and belief we need to at the practicalities of records management in these different ages.

2.2 The registry age

In the days before e-mail a gap in time and space existed between post arriving into an organisation and post arriving at the desk of the officer/official to whom it was intended. Organisations used this gap in time and space to interpose records clerks, organised in registries, to file documents and correspondence needed as records.

Post room staff would filter the morning correspondence:

- they would send business correspondence to the records clerks in the registries
- they would send post that was obviously personal or promotional in nature direct to the addressee.

The files created and maintained by the registries were relied on by all stakeholders. This meant that omissions in the file would be noticed by the only people in a position to notice them - the people carrying out that work.

2.3 The disruption of e-mail

The coming of e-mail created a rival communications channel to that provided by the postal system. Organisations had no time to plan how to deploy e-mail in a way that would enable them to transparently filter and file correspondence. The network effect meant that as soon as their customers and stakeholders adopted e-mail, then the organisation had to adopt it too. That meant deploying off-the-shelf commercial e-mail packages.

The introduction of e-mail had three major effects:

- it collapsed the gap in time and space between the sender and recipient of a document/communication

- it exponentially increased the volume of correspondence
- it gave individuals a new source of reference, their e-mail account, which meant they had less need to consult the official paper file, less occasion to notice any omissions on that file, and less reason to take action if there were gaps in the file.

2.4 The age of the electronic records management system

The introduction of e-mail meant that organisations lost control of their main communications channel. They tried to regain that control by requiring employees to move documents and correspondence needed as a record into an official file in an electronic records management system.

The organisation would provide employees with a generic definition of what constitutes a record, and expect them all to apply this definition to their own e-mail account.

However in practice each individual interpreted that definition differently. The amount of correspondence saved into the system depended on the motivation, awareness and workload of each individual.

Correspondence and documentation built up outside the electronic records management system, and outside of the protection of records retention rules. In particular correspondence built up in e-mail accounts.

Organisations faced a double edged sword:

- if they deleted e-mail accounts promptly they wiped out their own memory, because they were not capturing sufficient e-mails in their electronic records management systems BUT
- if they let e-mail accounts build up they were amassing huge quantities of trivial, private, and personal e-mails alongside the e-mails needed as a record.

Furthermore the one-to-one nature of e-mail communication lent itself to unguarded and sometimes toxic comments that also built up in e-mail accounts. This contrasted with the paper days when envelopes would be opened in the post room and correspondents would moderate their communications in the knowledge that their communications could be read by many different eyes on their way to the addressee.

2.5 The age of automation

An early step in the move to automation was the decision of the US Securities and Exchange Commission (SEC) (2) to require organisations engaged in the trading of financial securities to capture all communications made and received by its traders. Barclay T. Blair (3) has said that this ruling 'singlehandedly created the e-mail archive industry'. In a situation such as the trading floor it would be ridiculous to expect a trader engaged in some kind of misdemeanor to voluntarily declare into a record system the e-mails they used to inform their collaborators of their insider information. The only way to ensure accountability was to capture everything, including trivial and personal correspondence.

Another milestone in the march of automation was the release of SharePoint 2007 in late 2006. The records management model in SharePoint had individual knowledge workers simply able to right click on a document and select the option 'send to record centre'. Administrators were expected to configure rules to enable the SharePoint records centre to organise the documents that were

sent to it. Lying behind this model was the belief that knowledge workers had better things to do with their time than engage with the type of corporate records classification that had been the organising principle behind the electronic records management systems whose market share SharePoint eroded.

The most significant step in the march of automation was the passing of the Managing Government Records Directive (4) in the US which mandated the US National Archives and Records Administration to explore ways of automating records management.

In the context of records management NARA defined automation as any move to reduce the burden of records management on end users, by no longer requiring them to take a decision on every single document or e-mail they create or receive. This was the first time that the move to automation had come from within the recordkeeping professions themselves.

3. THE RANGE OF APPROACHES TO AUTOMATION

3.1 NARA's typology of approaches to automation

NARA released a report (5) in 2014 which listed the different ways in which records management might be automated.

These approaches can be grouped in two different categories, depending on the way in which they reduces the burden on end-users.

The first category of approaches continues to apply records management disciplines at the document/e-mail message level, but uses a machine rather than humans to determine what needs to be captured as a record and where it needs to be filed/classified. These approaches work by either:

- Defining workflows which automatically capture records at different stages of a process
- using auto-classification (through a machine learning tool or through the definition of rules) to select documents/e-mails needed as a record and to file them OR

The second category abandons the attempt to manage records at the document/e-mail message level, and instead manages records at a higher level of aggregation. These approaches involve:

- applying defensible disposition policies to existing groups of records (for example NARA's acceptance that Federal Agencies could preserve entire e-mail accounts rather than require individuals to select which e-mails met the definition of a record) OR
- holding a records classifications and retention rules in one application, and applying them to objects in the many different systems of the organisation (shared drive, SharePoint, Exchange etc.)

3.2 Vendor support for automation

All of these approaches are feasible, and supported by mainstream tools.

- Enterprise content management (ECM) vendors such as Oracle and Documentum and IBM have long provided sophisticated workflow definition tools with their products.
- ECM vendors such as Open Text and IBM provide auto-classification capabilities as part of their enterprise content management (ECM) suite.
- Content analytics tools and e-Discovery tools (such as Nuix, HP Control Point and others) give administrators a dashboard by which they can define parameters for particular types of content (for examples documents on a shared drive that are more than seven years old) and trigger a workflow to get the content within those parameters reviewed by the content owners and then destroyed if the content owners authorise the disposal
- In-place records management tools such as those offered by RSD and IBM enable an organisation to intervene in systems such as SharePoint and MS Exchange and link objects in those applications (libraries, content types or folders in SharePoint, folders in individual e-mail accounts) to the organisation's record classification and associated retention rules

4. EVALUATING APPROACHES TO AUTOMATION

4.1 Workflow

Of the above approaches the workflow approach is the one that most approaches the standard of reliability, comprehensiveness and usability achieved by the registry system approach in the paper age.

For example an insurance company might set up a workflow system for dealing with claims. They might ensure that communications related to claims are channeled into mailboxes specifically created for claims correspondence. They would use configure workflows to allocate a claims number to each claim, and to ensure that any subsequent correspondence relating to that claim and any recordings of voice conversations are kept together on one claim file.

Note how the insurance company has wrested control over the communications channel between claimants and its staff away from individual e-mail accounts and into mailboxes governed by its workflow system.

The biggest difference between the workflow approach in the digital age and the registry approach in the paper age is that:

- registry systems in the paper age could scale across all the different activities of an organisation (simply by maintaining an adequate ratio of records clerks to total numbers of staff) BUT
- the definition of workflows is too time intensive to enable an organisation to extend workflow control over the full range of its activities. Unless a work process is relatively predictable and often repeated, then there will be insufficient return on investment to justify defining a set of workflows to control the process.

4.2 Auto-classification

There are two methods of auto-classification. The first is by machine learning, Machine learning uses sophisticated statistical

algorithms to identify patterns within a particular set of documents. It works as follows:

- the administrator gathers together a sufficiently large sample set of document/e-mails that correspond to a particular category in a records classification
- the sample document set is fed to the machine learning tool
- the tool identifies the common patterns present in each document within the set, and is then able to go out and identify other documents/e-mails that correspond to that category

The second kind of auto classification is by the definition of rules. For example a rule might read: 'if a six digit project code appears in the subject line or text of the e-mail then move the record to the file that corresponds to that project code'.

The rule definition approach has:

- the advantage of transparency. It is easier for colleagues to trust an auto-classification tool if you can explain to them precisely the logic by which the tool will work. It is easier to explain the rules that have been written, than it is to explain the complex mathematics behind the machine learning tools.
- the disadvantage that it is even more time consuming to define rules for auto-classification than it is to build document sets to define a machine learning tool.

Both forms of auto classification share a common disadvantage.

- Records classifications tend to be very granular.
- The more granular a classification is, the more nodes it has at the bottom level
- The more bottom level nodes it has the more training sets that have to be gathered for the machine learning tool, or the more rules that have to be defined for the rules engine.

Another complication with records management is that we do not normally apply our classifications directly to documents. Instead we apply it indirectly, via containers/aggregations/folders/files that represent specific pieces of work.

These pieces of work emerge as new projects emerge. Ideally we need an auto classification tool to both

- identify which classification a document belongs to, for example whether it arises from an engineering project, a consultation, or from the management of a member of staff AND
- recognise specifically which engineering, project, which consultation and which member of staff it relates to.

In practice if you are going to try to apply auto-classification at a corporate scale, across a wide range of different activities, then you will end up using a 'big bucket' approach which will group records into mega-containers such as 'Environmental policy records' 'Health and Safety records', rather than granular containers such as 'Wind turbine policy 2012 to 2015' or 'Asbestos records for the HQ building'.

The problem is that it is hard to apply an accurate retention rules to big bucket containers. For example if you wanted to apply:

- an engineering retention rule that is triggered by the end of the life of a structure,

- a staff management retention rule that is triggered by the date the person left employment
- a consultation retention rule that is triggered by the closure of the consultation,

then you have to group together records into containers specific to one member of staff, one consultation, and one engineering project.

4.3 Defensible disposition

Defensible disposition is the least intrusive method of automation because in theory it involves no change to the way that content accumulates in an organisation. It simply gives you the tools to apply disposition rules to those accumulations.

The disadvantage of the approach is that some accumulations of content, most notably e-mail accounts, involve such a mixture of the trivial and significant; harmless and toxic; and the personal and business; that applying a retention rule to such an aggregation may involve unacceptable compromises.

4.4 In place records management

In place records management enables an organisation to maintain very sophisticated records classification and retention rules and apply them in different environments. It is most effective in organisations that are global in scope. Such organisations may need to apply different retention rules to records arising from the same function or activity in different jurisdictions. They are also likely to have a great many different content management systems.

In-place records management approaches intervene when particular events happen, for example when a new object such as a folder or a library or a site is created in SharePoint, or a new folder is created in an e-mail account. The intervention serves to link the object and its contents to the organisation's record classification and retention rules.

It is at its best when working with systems such as SharePoint and ECM systems that have a sophisticated API and a reasonable level of existing organisation. It is less effective with:

- e-mail accounts (if an individual does not use folders in their e-mail account then there may be a paucity of events to trigger the tool to intervene)
- shared drives (which lack an API to enable the tool to intervene properly).

5. INTERVENING IN THE E-MAIL COMMUNICATIONS CHANNEL

One of the weaknesses of the records management situation in organisations is that content tends to build up in ever larger accumulations. Cassie Findlay pointed out in a recent lecture that this puts records at risks, because the accumulations are so large that eventually sweeping decisions have to be made that affect all content within the aggregation.

The most glaring examples of this comes with e-mail accounts that organisations end up applying entirely arbitrary disposition rules to.

At the time of writing e-mail is still the main channel of communications into and out of most organisations. This situation will not persist for ever, but whilst it does persist it is important that we find a way to filter and control accumulations of

e-mail. When we look back at records management history we can see that the times in which we have been able to control the channels by which recorded information is communicated, then we have been able to build reliable and scaleable records systems.

We have seen that none of the proposed approaches to automation can yet deal satisfactorily with e-mail:

- NARA's Capstone approach to preserving some e-mail accounts permanently only helps historians in the distant future
- auto-classification only classifies into big buckets (when applied at corporate scale)
- defensible disposition approaches struggle to find a defensible retention period for e-mail accounts
- workflow can only stretch to a small number of processes.
- in-place approaches struggle if individuals do not use folders in their e-mail accounts

The continued failure of vendor offerings to help an organisation manage their e-mail should not stop organisations trying to win back control of the way that e-mails accumulate. In this section we explore two different ways that a manageable correspondence file could be filtered from individual e-mail accounts.

5.1 Treating e-mail accounts as correspondence files

In theory an e-mail account is simply the electronic equivalent of the correspondence files that many individuals kept in the hard copy age.

The two differences are that:

- in the hard copy age when an individual changed job within an organisation they left any correspondence files behind them for their successor. In the e-mail age most organisations allow an individual to keep that correspondence in their in-box even when they move to a completely different role
- in the hard copy age private and personal correspondence did not find its way onto a correspondence file, but in the e-mail age private correspondence accumulates cheek by jowl with business correspondence in the same account.

We have seen the relative failure of attempts to get individuals to filter their e-mail accounts into filing structures within electronic records management systems, due partly to the high volume of e-mails such individuals create and receive.

One response to this would be to pull back from the insistence on filing e-mails into filing structures, and instead create accumulations of e-mails that are non-toxic and which can be passed onto a post holders successors.

5.2 Role based e-mail correspondence files

One relatively simple way of filtering e-mail would be to:

- create a correspondence file for each role in the organisation
- link each individual's e-mail account to the correspondence file for the role they occupy
- intervene whenever an individual leaves a post. The purpose of the intervention should be to capture into the relevant correspondence file, all the e-mail from the time period in

which that individual spent in that post, minus any e-mails the individuals has flagged up as private.

- repeat the process when the new incumbent to the role leaves. The correspondence file would build up as different post holders occupied and then left the role.

The organisation would need to educate individuals that their e-mail will be passed on to their successor, and would need to give them a means of flagging e-mails that are private and should not be passed onto their successor

This approach creates a partially multi-faceted record. It extends access to the accumulation of e-mails to an individual's successor in post and their line manager. It could be used for compliance purposes by legal counsel.

5.3 Team based e-mail correspondence files

The United Nations Food and Agriculture Organisation (FAO) went one further than the role based correspondence file (6)

They intervened in the process whereby individuals send e-mails. When an individual presses 'send' in an FAO e-mail account they are faced with a pop-up

- The pop-up asks the sender whether or not the e-mail they are about to send is a record.
- If they select that it is a record, then a copy of the e-mail is routed to a record repository
- In the record repository the e-mail is stored in a correspondence file for the team with which that individual works
- Each team correspondence file is configured to send a digest e-mail to each member of the team once a day, listing all the e-mails tagged as 'record' by their team mates the previous day.

FAO found that some teams significantly reduced the number of e-mails that they copied to each other, because they know that by saving an e-mail as a record all their colleagues would become aware of its existence the following day via the digest e-mail.

What is interesting about the FAO approach is that they have paid as much attention to ensuring that the records are actually used and read as they have to ensuring that records are captured

In effect their record system was providing a current awareness tool for colleagues, who could see what their colleagues were working on without the intrusion of being copied into multiple e-mails.

6. CONCLUSION

The current array of automated approaches present us with a dilemma.

- The approach that is the most effective (the workflow approach) is not scaleable across all of an organisations activities because of the time and resource necessary to analyse processes in order to build the workflows
- The approach that is the most scaleable (the auto-classification by machine learning) achieves that scaleability at the expense of a loss of granularity that would see records grouped into 'buckets' that are simply too large to enable us to apply useful retention and access rules

In-place records management tools and content analytic tools are pragmatic approaches to the messy situation that organisations find themselves in, with content scattered over many different repositories. However neither one of these two tool sets has yet to provide a solution to the problem of the build up of large and unmanageable e-mail aggregations.

The best records management examples we have looked at in this paper were the following:

- the registry systems operated in the paper age
- line-of-business workflow/case-file systems such as the insurance claims systems
- the FAO's use of team based e-mail correspondence files

These three approaches each had two things in common:

- they each routinely intervened in the communications channel by which individual colleagues sent and received written/recorded communications
- They each captured records that were read and relied upon by the officers/officials carrying out the work that the records arose from

The point about intervening in the communications channel is important. Without that intervention the record system is 'outside of the loop', an after-thought, sitting to one side of the way business is conducted rather than engrained in the way that business is conducted.

Any system will develop imperfections. Sustainable systems have built in provision for spotting imperfections and correcting them. Records systems must be referred to by end users in order to be sustainable - this is because the colleagues carrying out a piece of work are the only people in the organisation who are in a position to notice gaps in the record and to do something about those gaps.

The challenge for automated approaches is this - how do you reduce the burden and responsibility of records management on end-users, whilst still retaining an active role for the end-user in the records system, as record users, and as whistle-blowers on gaps and imperfections?

7. REFERENCES

- [1] NARA's Capstone Bulletin
<http://blogs.archives.gov/records-express/2013/08/29/capstone-bulletin-issued/>
- [2] SEC rule 17a-http://en.wikipedia.org/wiki/SEC_Rule_17a-4
- [3] Barclay T Blair's quote is in
<http://barclaytblair.com/2013/06/21/response-to-naras-capstone-email-bulletin/>
- [4] Managing Government Records Directive
<http://www.archives.gov/records-mgmt/prmd.html>
- [5] NARA report into automation
http://blogs.archives.gov/records-express/files/2014/03/Automated-Electronic-Records-Management-Report-and-Plan_3.6.14_finaldraft.pdf
- [6] Case study of the FAO's records system implementation
<http://thinkingrecords.co.uk/2013/07/13/faos-approach-to-making-e-mail-manageable-and-shareable/>