The National Archives

The application of technology-assisted review to born-digital records transfer, Inquiries and beyond

Research report

Published: February 2016

Contents

Glossary of Terms

Algorithm	A self-contained step-by-step set of operations to be performed to solve a specific problem. Algorithms exist to perform calculations, data processing and automated reasoning.
Appraisal and selection	The process of distinguishing records of continuing value from those with no further value.
Born-digital records	Records created originally in digital formats – such as emails, documents, spreadsheets – as opposed to paper records that have been digitised.
Data analytics software	Generic term for software that can index data sets and extract patterns and connections.
Digital sensitivity review	The process of identifying sensitive content in digital records that should be exempt from release.
eDisclosure	The phrase used for the eDiscovery process in the United Kingdom.
eDiscovery	The discovery or disclosure of electronic information for the purposes of litigation. This phrase is used in the United States but is also the common descriptor for software tools that assist with eDiscovery/eDisclosure in the United Kingdom.
eDiscovery software	Software that can index data sets and extract patterns and connections. Usually associated with the legal discovery process during litigation.
Electronic Discovery Reference Model	A model designed to represent the eDiscovery process (see EDRM.net).
Electronically stored information	Information created, managed and consumed in digital form, which requires the use of computer hardware and software to access it.
F-measure	The harmonic mean between precision and recall.
Latent Dirichlet Allocation	A topic modelling algorithm that automatically detects groups of topics from the content.
Latent Semantic Indexing	The singular value decomposition mathematical technique used to detect terms and concepts and find patterns and relationships.
Predictive coding	A way of automatically classifying documents based on statistical analysis and machine learning. Computers are ‘taught’ to identify patterns and this iterative training can enhance the accuracy of results. ‘Predictive coding’ can also be used to mean technology-assisted review.
Regular expression	A type of advanced search pattern where the user specifies the structure of the expression to be found rather than a search on the specific letters or numbers to be found. It could be used, for example, to find email addresses, credit card numbers, UK NI numbers, etc. Used where the actual names and numbers are not known but the structure of the expression can be predicted.

Relational databases	A database that recognises relationships between common data fields.
Technology/Computer-assisted review	A process involving expert document reviewers using a combination of computer software and tools to electronically classify records.

Executive Summary

Born-digital records pose many challenges for government departments, including high volumes of records and a lack of structure in born-digital record collections. These affect not just the appraisal, selection and sensitivity review processes when transferring records to The National Archives, but also pose challenges for departments responding to Inquiries and Freedom of Information requests. Additionally, there are broader information management and security concerns for born-digital record collections.

To examine these challenges and explore potential solutions, The National Archives conducted trials of eDiscovery software and looked at additional research to test how these tools and processes could help meet the challenges of born-digital records. This report summarises the key lessons learned from that work.

The report concludes that technology-assisted review using eDiscovery software can support government departments during appraisal, selection and sensitivity review as part of a born-digital records transfer to The National Archives. This support also extends to responding to Inquiries and Freedom of Information requests, as well as information management and information security. We summarised these findings into eight lessons learned:

1. Understanding born-digital collections at a high level
2. Reducing the amount of information to review
3. Extracting meaning
4. Identifying personal information
5. Procurement
6. User interface
7. Collaboration with other teams
8. Confidence in technology-assisted review

There is no completely automated solution; human input is still required at all stages. However, technology-assisted review offers ways to understand, value and prioritise born-digital records, as well as reducing the volume needing to be manually reviewed. The report ends by setting out further research The National Archives will conduct and the support it plans to give to government departments to help them manage their born-digital record collections. As such, we will continue to work with the Cabinet Office and Government Digital Service. In addition, we will continue collaborating with other centres of expertise within government and beyond to enhance methods and tools.

1. Introduction

From 2016, born-digital records[1] will start to make up a growing proportion of the information transferred from government departments to The National Archives. The transition from the 30-year rule to the 20-year rule under the Public Records Act is accelerating this process. Understanding how this transition will affect existing processes has been part of The National Archives' Digital Transfer Project.

There are a number of challenges and concerns when transferring born-digital records. The main concerns are an increase in volume and potentially ephemera, along with less structure and diminished context compared with paper records. This is making appraisal, selection[2] and sensitivity review more difficult when transferring records to The National Archives. In the digital transfer process, these stages relate to Gateways 1 and 2.

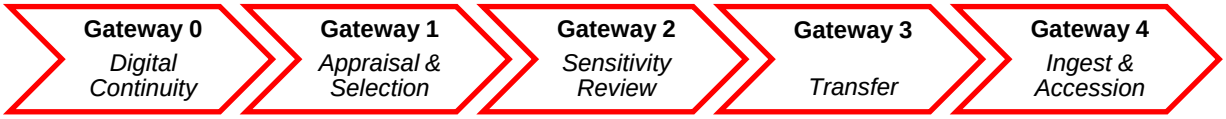

Figure 1: Five Gateways of the digital transfer process

These concerns have acted as a driver for The National Archives to conduct research and trial existing software tools that could help government departments to address these challenges. Born-digital records are also more generally affecting government departments' information management capabilities, with specific implications for departments responding to Inquiries and Freedom of Information requests.[3] This report summarises the lessons learned from the software trials and additional research, which were conducted in 2015.

The research focused on eDiscovery processes and tools and the applicability of technology-assisted review for appraising, selecting and sensitivity reviewing born-digital records. eDiscovery is a process concerning the discovery or disclosure of electronic information for the purposes of litigation.[4] eDiscovery practitioners have modelled processes and software tools have been developed to support the activity. This includes reducing the volume of digital information to review and extracting the most relevant material. Given the similarities between the activities involved in eDiscovery and the problems posed by born-digital record collections, it was these processes and tools that The National Archives evaluated.

Technology-assisted review is a combination of input from expert human reviewers and computer software to partially automate the classification of records and to

[1] Born-digital records are those created originally in digital formats – such as emails, documents, spreadsheets, videos and images – as opposed to paper records or photographs that have been digitised.
[2] Appraisal and selection is the process of assessing and choosing the records that should be kept by departments for continued business use, preserved because they contain ongoing historical value and sent to The National Archives or destroyed because they have no further value. For born-digital records, this is conducted at the highest level possible (i.e. at a macro level – for example, at a business function or series level) given the high volumes and the need to conduct the process efficiently.
[3] For the purposes of this report, the term 'Inquiry' relates to an official review of events or actions ordered by the government.
[4] eDiscovery is the more popular term used in the United States although the process is known as eDisclosure in the United Kingdom. However, most software packages in this field are branded eDiscovery tools and we therefore adopt that term in this report.

identify patterns and similar content.[5] This report provides insight into the possible application of eDiscovery software for technology-assisted review to tackle the challenges of born-digital records. It also highlights further research to be conducted by The National Archives to build on these findings. This report is written as a practical first step in an ongoing process to refine the support and guidance for the management and transfer of born-digital records.

2. Technology-assisted versus manual review processes

To understand the challenges that born-digital records pose, The National Archives consulted with government departments, international organisations, academics and members of the Advisory Council on National Records and Archives. This formed part of a wider piece of work by The National Archives to document the digital landscape in government.[6]

In addition, Sir Alex Allan's *Review of Government Digital Records* was published in December 2015.[7] This summarised a number of challenges and recommendations concerning the handling of born-digital records, including:

- identifying the best technologies to manage digital information
- finding software tools to help organise and search legacy digital data
- The National Archives and Government Digital Service to lead on a centralised strategy for legacy digital records, so that departments do not seek independent solutions
- more research on sensitivity review to be undertaken, led by The National Archives with support from the Government Digital Service and drawing on academic research as appropriate
- ensuring sufficient high-level buy-in and collaboration between The National Archives and the Government Digital Service.

There is a general acceptance that processes designed for the review of paper records collections will not meet the challenges of born-digital records. Applying manual review processes for large volumes of born digital records paired with the issue of declining resources mean that existing methods need to be adapted. It would be easy to think that technology-assisted review is an inferior process to manual review and only employed because of the challenges of digital information. However, there is evidence that technology-assisted review can be as accurate, if not more accurate, than manual research or keyword searches alone.[8] Although

[5] Technology-assisted review is also known as computer-assisted review but the former term is used in this report for consistency.
[6] The National Archives (2016) 'The Digital Landscape in Government 2014-2015. Business Intelligence Review', nationalarchives.gov.uk/information-management/manage-information/our-research/.
[7] Allan, A., (2015) 'Review of Government Digital Records', *Cabinet Office,* gov.uk/government/publications/government-digital-records-and-archives-review-by-sir-alex-allan.
[8] Grossman, M., and Cormack, G., (2011) 'Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review', *Richmond Journal of Law and Technology,* Volume 17, Issue 3, p.33.
Oard, D., Baron, J., Hedlin, B., Lewis, D., and Tomlinson, S., (2010) 'Evaluation of information retrieval for E-discovery', Artificial *Intelligence and Law*, Volume 18, Issue 4, p.7, http://terpconnect.umd.edu/~oard/pdf/jail10.pdf.

many people may consider human review to be the ‘gold standard’ of review, in reality this may not be the case.

For example, research by Blair and Maron in 1985, and replicated later, showed that keyword searches returned only 20% of the relevant documents, while searchers using technology-assisted review estimated they had found at least 75%.[9] Research has also shown that readers follow an F-shaped pattern when viewing documents, meaning they prioritise certain spaces on a page and miss critical content in other areas. These findings were revealed by research that tracked users as they read from a screen and then heat maps were constructed to illustrate their reading patterns. This means that human reviewers are likely to miss information, especially if it is outside the ‘F-shape’.

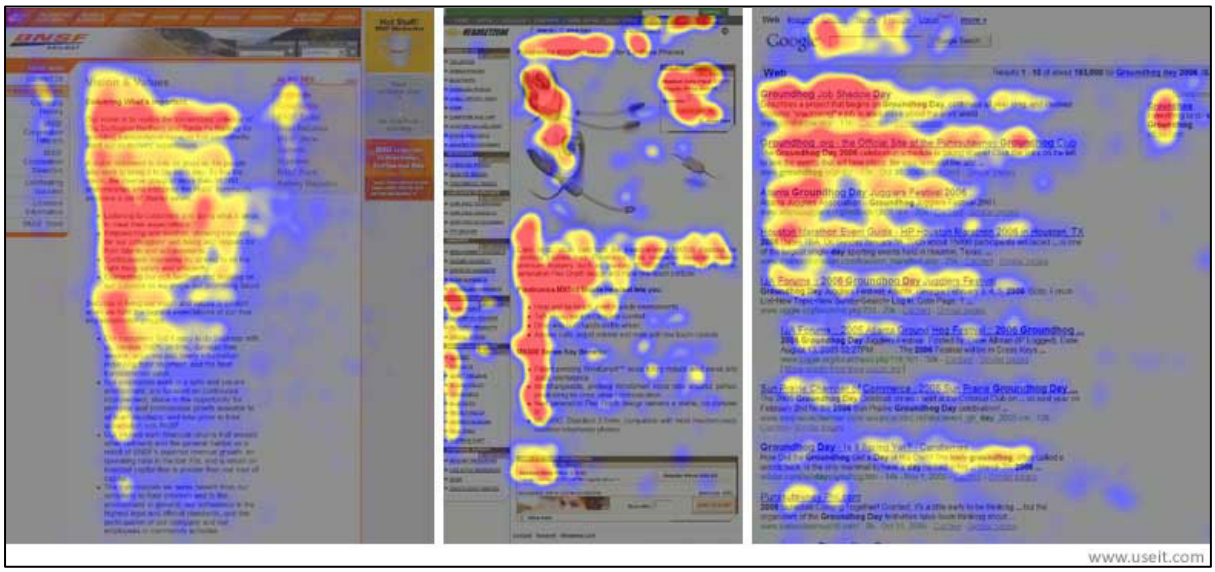

Figure 2: F-Shaped Pattern for Reading Web Content[10]

In contrast to human review, technology-assisted review using software tools can index and search *all* content and metadata equally. Technology-assisted review using these software tools can prove more powerful than manual review and keyword search alone, while at the same time helping reviewers handle large volumes of information.

The eDiscovery tools that can facilitate technology-assisted review are not new and have been deployed in the legal field for a number of years. They use predictive coding (which is a way of automatically classifying documents based on statistical analysis and machine learning) to identify documents that are relevant to the reviewer. It involves a learning process, which requires the reviewer to identify a relevant subset of information from a larger collection to train the software. Algorithms in the software then use this ‘seed set’ to find conceptually similar

---

Peck, A., (2011) ‘Search, Forward. Will manual document review and keyword searches be replaced by computer-assisted coding?’, *Law Technology News,* https://openairblog.files.wordpress.com/2011/11/peck-search-forward.pdf.

[9] The Blair and Maron evidence is cited in: Peck, A., (2011) ‘Search, Forward. Will manual document review and keyword searches be replaced by computer-assisted coding?’, *Law Technology News,* p.1, https://openairblog.files.wordpress.com/2011/11/peck-search-forward.pdf.

[10] Nielsen, J., (2006) ‘F-Shaped Pattern For Reading Web Content’**,** www.nngroup.com/articles/f-shaped-pattern-reading-web-content.

information in the larger collection.[11] The strength of these technologies primarily relies on searching text; therefore they will be less useful where the information is not text-based unless they offer additional functionality (such as skin tone recognition in photographs). They also require iteration, with the reviewer refining and correcting the results until the software is returning a level of accuracy that they find acceptable (organisations will need to define their own acceptance criteria based on their risk appetite).

Use of these technologies and predictive coding is becoming increasingly accepted. For example, in February 2012 Magistrate Judge Andrew J. Peck of the United States District Court issued an opinion approving the use of technology-assisted review as 'an acceptable way to search for relevant' electronically stored information.[12] In 2015, the High Court in the Republic of Ireland endorsed the use of predictive coding for an eDiscovery exercise in the case of the *Irish Bank Resolution Corporation Limited v Sean Quinn.*[13] These examples are an indication of the potential direction of travel in the United Kingdom towards the wider use of technology-assisted review. Outside the legal profession, predictive coding is also used in information security where it forms the basis of spam filters.

These findings led The National Archives to examine the eDiscovery market and consider whether these tools, which are good enough for use in courts, are good enough for records managers. The main focus of the research was on three software trials. This was informed and validated by speaking to United Kingdom (UK) government departments using or procuring eDiscovery tools. We also consulted with international eDiscovery experts such as Jason R Baron of Drinker Biddle & Reath LLP and former Director of Litigation at the National Archives and Records Administration in the United States. This was supplemented with reviews of academic literature and an Information Management Liaison Group meeting on interrogating born-digital records collections held at The National Archives in November 2015.

3. Objectives of the software trials

3.1. Overall

The overarching objective for the software trials was to understand the extent to which existing software tools can assist government departments in increasing the efficiency of appraising, selecting and sensitivity reviewing born-digital records. Because the objectives and methods of the legal process of eDiscovery broadly align with the objectives and processes of Gateways 1 and 2 of the digital transfer process, it was decided to explore software in the eDiscovery market.

3.2. Gateway 1 - Appraisal and selection

The software trials looked at how technology could help departments extract meaning from their born-digital collections. In particular, functionality that could

[11] Hampton, W., (2014) 'Predictive Coding: It's Here to Stay', *E-Discovery Bulletin,* www.skadden.com/sites/default/files/publications/LIT_JuneJuly14_EDiscoveryBulletin.pdf.
[12] Grossman, M., and Cormack, G., (2013) 'Glossary of Technology-assisted Review', *Federal Courts Law Review,* Volume 7, Issue 1, p.13, www.fclr.org/fclr/articles/html/2010/grossman.pdf.
[13] High Court of Ireland, (2015) 'Irish Bank Resolution Corporation Ltd versus Quinn', IEHC 175, www.bailii.org/ie/cases/IEHC/2015/H175.html.

automatically categorise information in unstructured born-digital record environments (i.e. collections of born-digital records with no applied classification scheme or folder structure) was tested to understand if it could support macro-appraisal and selection decisions at scale.

Shared drives are often a source of the majority of this unstructured information and their prevalence was revealed in a review of the digital landscape in UK government. The report concluded that two-thirds of born-digital information is stored in shared drives versus one third stored in electronic records management systems.[14]

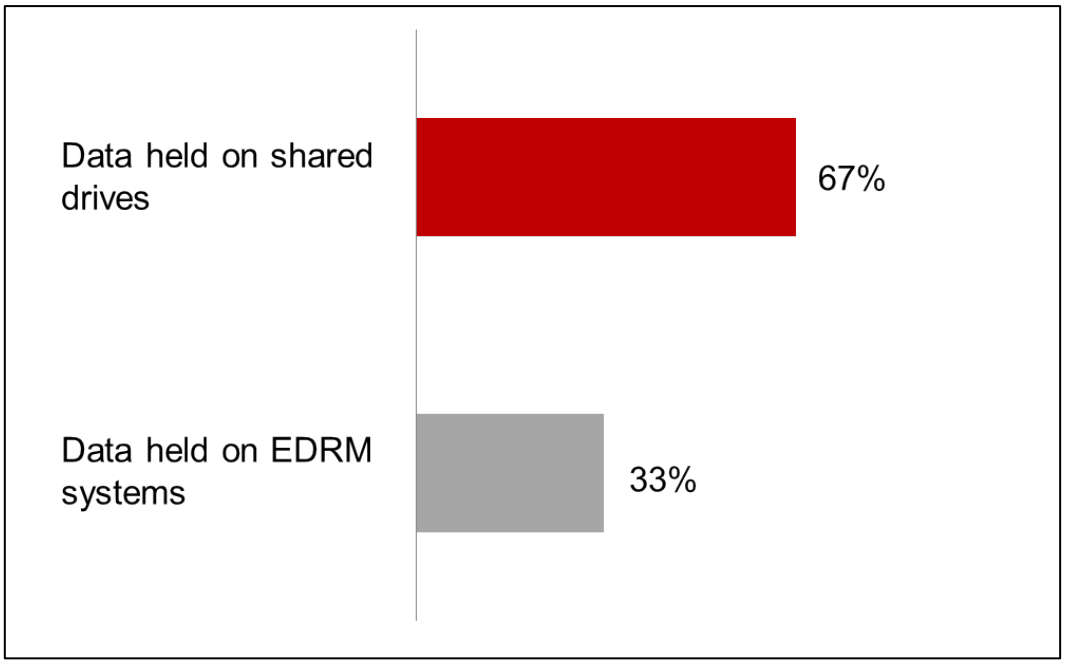

Figure 3: Percentage of data from 11 of the key 21 government departments held in EDRM systems vs shared drives[15]

3.3. Gateway 2 - Sensitivity review

Born-digital records present a challenge across a breadth of sensitivities. This includes exemptions relating to national security and international relations. However, because personal exemptions account for around 75% of exemption requests and it was assumed that they could be more easily defined (i.e. their format and length can be predicted), it was decided the trial should focus on identifying personal information.[16] The remaining 25% of exemptions are more complex and will require further research.

[14] The National Archives (2016) 'The Digital Landscape in Government 2014-2015. Business Intelligence Review', p.21, nationalarchives.gov.uk/information-management/manage-information/our-research/.
[15] The key 21 government departments are the ones that The National Archives have identified as accounting for 90% of the records transfers to The National Archives over the last three years.
[16] Personal exemptions have been classified as Freedom of Information Act exemptions 38, 40(2) and 41. The 75% statistic is based upon research by The National Archives, which looked at the exemptions for 21 government departments. The National Archives (2016) 'The Digital Landscape in Government 2014-2015. Business Intelligence Review', p.26, nationalarchives.gov.uk/information-management/manage-information/our-research/.

3.4. Automated redaction

A requirement associated with the identification of sensitive personal information is the ability to redact sensitive content from documents before they are transferred to The National Archives and made available to the general public. The software trials, therefore, also investigated whether the tools could facilitate the redaction of sensitive information in a scalable, reliable, consistent and defensible way.

3.5. User experience

Another set of issues explored during the trials concerned how the outcome and process of this technology-assisted review compared with a manual process. This included assessing how user-friendly the tool was by looking at factors such as how much training a reviewer might need to use it confidently and whether information was presented in a helpful user interface.

Furthermore, we looked at how the tools could support workflows associated with appraisal, selection and sensitivity review as well as how the reliability of the results compared to a manual approach.

3.6. Other applications

In addition to the specific application of eDiscovery tools to support the digital transfer process, the software trials were an opportunity to consider broader uses. This included other processes and challenges that share similarities, such as responding to Freedom of Information requests and requests from Inquiries. Furthermore, information management, digital continuity and information security challenges more broadly could, in part, be supported by eDiscovery software (e.g. to understand the breadth and depth of digital collections and to identify valuable and sensitive data that needs particular levels of information security protection). In this respect, the tools could have application across the information management spectrum and enable an integrated approach across data management, security and technology.

3.7. Out of scope and limits to the study

Not in the scope of the trials were assessments of hardware or software performance, which can be device and network dependent. Conclusions on the scale of volumes that can be handled are limited to the scale of the born-digital records sample we used (circa 100,000 records). The cost, procurement methods and ease of deploying the software on existing information technology infrastructures in government was also out of scope.

4. Methodology of the software trials

4.1. Software requirements

The intended outcome of the trials was to learn about technologies and useful features, and not to choose or recommend a specific software tool or supplier. To that end, the identities of the software tools and suppliers have not been disclosed in this report. There are already a number of active players in the eDiscovery field with

a range of software tools with functionalities that can meet the requirements for understanding information collections at a high level.

The requirements for the trial were derived by breaking down existing appraisal, selection and sensitivity review processes into their component parts. Typically, these high-level requirements can be met by eDiscovery software. This is, in part, because the eDiscovery process follows a similar set of stages to appraisal, selection and sensitivity review. In the legal field the process has been mapped as the Electronic Discovery Reference Model.[17]

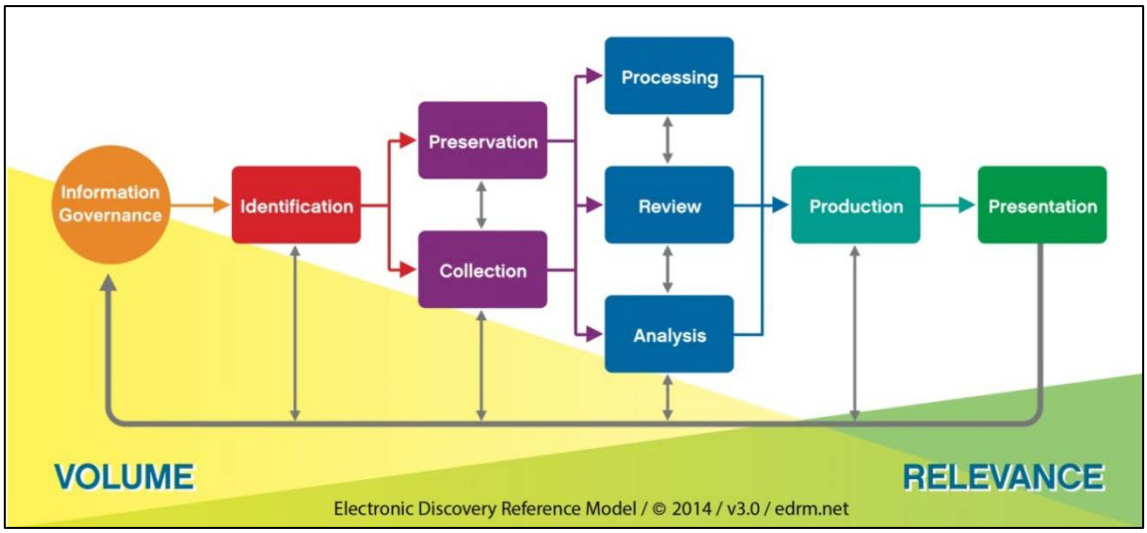

Figure 4: The Electronic Discovery Reference Model

Note that in the eDiscovery model the volume of information (the yellow triangle) reduces throughout the process while the relevance increases (the green triangle). This is achieved by removing less relevant information and leaving only the more relevant information.

For the software trials, the tools were expected to support the following requirements, which align with the eDiscovery model:

- **identification, preservation and collection** – the ability to identify and index a large collection of digital information with multiple file formats
- **processing** – the ability to preserve metadata, normalise formats and reduce data volumes for review
- **review and analysis** – the ability to gain an understanding of document content and organise the information into logical categories.

Underpinning these requirements was a fourth enabling one:

- **ease of use** – the ability to support simple workflow processes and provide an intuitive user interface experience.

[17] The Electronic Discovery Reference Model is available online at: www.edrm.net/resources/edrm-stages-explained.

Many eDiscovery software tools would be able to meet these requirements and facilitate technology-assisted review through predictive coding. However, predictive coding covers a number of different techniques, meaning the tools have different algorithms and methodologies.[18] To capture some of this diversity, we chose three tools that were representative of different technologies including:

- Latent Dirichlet Allocation, which is a topic modelling algorithm that automatically detects groups of topics from the content of documents
- Relational Databases, which recognise relationships between common data fields
- Latent Semantic Indexing, which uses the singular value decomposition mathematical technique to detect terms and concepts and find patterns and relationships.

Because of their functionality, it is worth noting that tools that share similar technology and meet similar requirements may already be available to some government departments. Legal teams and statisticians may already utilise tools to help them sort, search and analyse information. Investigative teams may also be consumers of these products. For example, eDiscovery is used to support digital forensic investigations.[19] These existing complementary requirements can strengthen business cases for justifying expenditure. Justification can also be enhanced by identifying additional use cases around information management including information security (e.g. identifying the truly sensitive information that must be protected with enhanced security) or delivering savings by reducing storage costs through deduplication.

4.2. Supplier briefing

The same briefing was shared with three software suppliers. The National Archives used the same corporate data, and interrogated the data the same way, using the same questions, in order to produce comparable results. Almost 100,000 born-digital records from The National Archives were used for the trials. A subset of almost 2,000 was manually sensitivity reviewed. We removed most of the filing structure in order to replicate an unstructured information collection. The key question was to understand how the tools could assist in increasing the efficiency of appraisal, selection and sensitivity review.

5. Lessons learned

This section of the report contains the key findings from the software trials. The lessons learned are presented within the context of the main challenges posed by born-digital records collections. The lessons are then presented to demonstrate how

[18] For an accessible high-level overview of the different methodologies and predictive coding more generally, see: Hampton, W., (2014) 'Predictive Coding: It's Here to Stay', *E-Discovery Bulletin,* p.29, www.skadden.com/sites/default/files/publications/LIT_JuneJuly14_EDiscoveryBulletin.pdf.

[19] See the Home Office report on eDiscovery and digital forensic investigation available at: gov.uk/government/uploads/system/uploads/attachment_data/file/394779/ediscovery-digital-forensic-investigations-3214.pdf.

the tools could help to meet the challenges in the context of activities including digital transfer to The National Archives and responding to government Inquiries.

5.1. Lesson learned 1: Understanding born-digital collections at a high level

What is the challenge?

Understanding the content of born-digital information collections can be challenging. There may be a number of information collections or data stores that are used by different business functions. These stores could include Electronic Document and Records Management or Enterprise Content Management systems, shared and personal drives, email accounts and other business tools (e.g. correspondence-handling software).

These born-digital information collections may or may not be structured and actively managed. They may be located within a department as a result of machinery of government changes and therefore there may not be existing corporate knowledge about the content and format of the information. This can prove a challenge when attempting to meet the recommendations in the Code of Practice on the management of records issued under section 46 of the Freedom of Information Act 2000. This, in turn, has implications for compliance with the Public Records Act and the ability to respond to government Inquiries.

What we learned

Overall, eDiscovery tools can assist organisations in understanding their digital information collections at a high level. For example, the software tools can provide insight into born-digital records collections by key criteria including:

- date
- format
- volume
- proportion of exact duplication in the collection.

High-level reports can then be routinely generated to monitor the overall digital collection, delivering benefits for information management, digital continuity and security more broadly. More specifically, these reports can be used to support activities such as appraisal and selection or a machinery of government change, requiring departments to identify born-digital records for transfer.

5.1.1. Gateway 0 – Digital Continuity

This high-level understanding of collections with the ability to generate reports will allow organisations to implement or improve their digital continuity plans – Gateway 0 in The National Archives' digital transfer process. Improving knowledge of new or existing information collections will enhance knowledge around risks and, where appropriate, enable Information Asset Registers to be created or updated.[20]

[20] The National Archives provides guidance on Digital Continuity and Information Asset Registers. For Digital Continuity see: nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/ and for Information Asset Registers see: nationalarchives.gov.uk/information-management/manage-information/planning/information-principles/information-valued-asset/.

Understanding file formats will also enable the organisation, if necessary, to plan for the migration of records that may otherwise become obsolete. This will ensure that material selected for preservation can be accessed until it is transferred to and accepted by The National Archives.

5.1.2. Gateway 1 – Appraisal and Selection

Understanding the age and volume of digital records will enable organisations to plan their born-digital records transfers to The National Archives in compliance with the Public Records Act. It will also help organisations comply with demands for information from Inquiries as well as being able to answer broader questions about their collections more easily and efficiently (e.g. 'do you have any material dating back to 2005?'), although the accuracy of the responses will still be dependent on the accuracy of the information they are recalling. If, for example, creation dates have been incorrectly updated during a data migration process, the tools may not be able to display the original creation date.

It will also help to identify events, people and places, which are important for selecting records based on their relevance and value. Once again the accuracy of the information, such as the correct spelling of names, will need to be accounted for. In general, however, the most efficient way of understanding the age and volume of digital records will be a technology-assisted review process based on human reviewers using efficient software tools to appraise and select the relevant born-digital records.

5.2. Lesson learned 2: Reducing the amount of information to review

What is the challenge?

The volume of digital information held by departments is likely to be large and split across a range of information stores (e.g. email accounts, records management systems and shared drives). It will contain varying degrees of ephemeral material and duplicate or near-duplicate information. The sheer volume of information makes digital continuity more challenging by increasing the task of knowing the collection and increasing storage costs. It is also a barrier to precise recall of information because it complicates the processes of searching for and retrieving the right information. At large scale, in particular, the reliance on keyword searches alone cannot guarantee the identification and extraction of all relevant content. This means that there are potential implications for complying with the Data Protection Act, Freedom of Information Act and responding to Inquiries if the methods to search and retrieve relevant information are inadequate (e.g. relying solely on keyword terms).

What we learned

The eDiscovery software tools are able to facilitate a 'funnel' approach to analysing born-digital records collections. The amount of information to review can be systematically reduced by using the tools to:

- identify the subsets of records to be reviewed (i.e. exclude records that are clearly identifiable as out of scope)

- identify duplicates that can be excluded from review, which proportionally can be very significant[21]
- identify near-duplicates that can be excluded from review (this can also be useful for finding drafts as opposed to final versions of documents)[22]
- systematically exclude file formats from the review (e.g. temporary files)
- focus on important formats (e.g. emails or photos), depending on the nature of the collection.

By using this 'funnel' approach, a large collection can be broken down into smaller subsets that are more manageable to review. It is important to note that this is not a fully machine-led approach. A reviewer will need to 'think' before they 'look' in order to use the software in an intelligent way based upon the specific characteristics of their collection. For example, in some departments it may be reasonable to exclude entire format types – such as images – from a collection because they are not considered records of long-term value or because they are not sensitive.

5.2.1. Gateway 1 – Appraisal and Selection and Gateway 2 – Sensitivity Review

There is no sole technological solution to appraisal, selection and sensitivity review; there will still be a need for human input. However, by using tools and adopting the 'funnel' approach, as well as prioritising review, the resources required to conduct Gateways 1 and 2 could be significantly reduced.

The process may also identify content that can be deleted (subject to internal retention and disposal schedules) or simply excluded from review, giving additional benefits for information system performance and data storage and migration costs. Similarly, the tools can be used to reduce the amount of information to sensitivity review.

5.2.2. Information security

Although information security was not originally a driver in the remit of the software trials, there were signs that the tools could support it. In addition to reducing the amount of information to review, the tools could be used to manage information collections based on risk. Collections of high-risk information containing sensitivities that need to be carefully managed for information security purposes may be identified. Separate policies and monitoring regimes could then be applied to this subset of information or it could be migrated to a more secure information store. This may allow for a more intelligent information management process that applies the appropriate level of information security to the relevant subsets of information rather

[21] A 2012 survey on digital information by Symantec reported that 42% of digital information held by organisations was duplicated (see: www.symantec.com/en/au/about/news/release/article.jsp?prid=20120703_01). Another survey by FindLaw quoted an industry average of 21% duplication (see: http://technology.findlaw.com/ediscovery-guide/processing-metrics.html).

[22] A 2012 report sponsored by an eDiscovery software company suggested 30-50% of electronic files in a court case are near-duplicates (see: Gunning, K., (2012) 'eDiscovery Document Review: Understanding the Key Differences between Conceptual Searching and Near Duplicate Grouping', www.equivio.com/files/files/White%20Paper%20-%20Understanding%20the%20Key%20Differences%20Between%20Conceptual%20Searching%20and%20Near%20Duplicate%20Grouping.pdf

than applying blanket policies. This may help to reduce costs through amended information storage planning. Furthermore, information risk could then be adequately identified and mitigated. The benefits of these software tools to wider information security concerns should therefore be investigated further.

5.3. Lesson learned 3: Extracting meaning

What is the challenge?

Born-digital records present a number of challenges for extracting meaning. In the majority of cases, it will not be feasible for reviewers to read all the documents or to know in advance what all the key events, themes and people are. Some born-digital records may be contemporary, some might be old (20 years old or more) and there could be poor contextual information.

This lack of understanding may be amplified by limited corporate knowledge of the subject or departmental function to which the records relate. This could be due to high staff turnover, resulting in the loss of corporate knowledge, or machinery of government changes that transfer new responsibilities into a department.

Overall, these issues of volume, context and staff knowledge will have the effect that the meaning, relevance and value of the born-digital records could be difficult to determine.

What we learned

It is possible to use eDiscovery tools to extract meaning from a large collection of born-digital records. In particular, the trials demonstrated the following three functionalities, which enable reviewers to extract meaning (it is important to note that different software providers use some of these terms to mean different things):

- **categorisation** – the tools can automatically group together conceptually similar born-digital records creating the ability to spot patterns within the information
- **clustering** – based on pre-existing, ‘man-made’ categories (e.g. departmental file plans) associated with selected examples (a ‘seed set’) of each category, some tools can group together born-digital records that are conceptually similar to the selected examples that the tool has been given
- **email visualisation** – some tools can help visualise and analyse email collections by showing the frequency of interactions between individuals over time.

These functionalities combined, or in isolation, allow meaning to be extracted from large digital collections even if there is little or no structure or pre-existing corporate knowledge about the information. In doing so, human input is enhanced (although not fully replaced) by this technology-assisted review. Academic research supports

this finding and even suggests that technology-assisted review can be superior to manual review.[23]

5.3.1. Gateway 1 - Appraisal and Selection

The tools can help organisations to understand their information collections at a high level. This is especially useful when information is stored in unstructured shared drives or inherited from machinery of government changes with limited prior corporate knowledge of the content. From this, topics and themes can be extracted to provide a high-level structure.

Categorising and clustering information can reveal conceptually similar information that can be subject to more consistent review and the application of macro-level appraisal and selection decisions. This means having identified topics or using existing structures the tools can assist in finding content similar to another document by subject or concept.

Email visualisation could be useful in revealing value for appraisal and selection decisions that otherwise may be invisible or harder to spot for the reviewer. The ability to identify key decision- makers from email exchanges may lead reviewers to identifying further leads to follow up. This functionality may be available in some products and not in others. Alternatively, they could be acquired as a standalone product (although none of these were tested as part of this software trial). Consideration over the necessity of this functionality should take into account the nature of the information collection (i.e. the volume and/or relevance of email).

5.3.2. Gateway 2 - Sensitivity Review

These functionalities could assist with identifying some sensitive subject areas within a born-digital record collection and lead reviewers to areas of possible sensitivities more quickly. Using the tools to find conceptually similar records across collections by extrapolating from examples of sensitive material introduces scalability to the digital sensitivity process by giving an alternative to file-by-file review. A risk-based approach can then be taken to determine whether to focus on or exclude particular information. Sampling can be used to validate those assumptions and determine whether the information is indeed sensitive or on the contrary it is not. To date, our analysis has predominantly focused on personal sensitivities. However, the possibilities referenced here mean that further research could be conducted on the effectiveness of eDiscovery tools for a wider range of sensitivities.

5.3.3. Inquiries

Responding to requests from Inquiries could be supported by the ability of eDiscovery tools to identify relevant documents more quickly and efficiently and find conceptually similar documents across the collection. This could provide content to disclose and/or reassurance that a nil or limited return can be justified.

The software tools offer better results than keyword search alone. Research has shown that there are significant limitations with keyword searches.[24] eDiscovery

[23] Grossman, M., and Cormack, G., (2011) 'Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review', *Richmond Journal of Law and Technology,* Volume 17, Issue 3.

tools do not require the end users to know the specific information being sought (as with a keyword search). Instead the tools can be trained with a 'seed set' of documents to automatically find similar content without relying on keywords.

The tools, if combined with an effective governance framework, can also provide a documented process for how a search for an Inquiry request was undertaken. This could prove necessary for justifying the scope of a search effort and for retrospectively applying the same processes to different information collections or stores. Such auditing will need to withstand sufficient scrutiny, which is why a well-established methodological technology-assisted review process using eDiscovery software tools can add weight to efforts to respond to Inquiries. Technology-assisted review is increasingly being accepted in legal cases, which require a high-level of rigour and audit. These processes and the authority they infer (not to mention the significant body of academic research that underpins it) can be adopted for the purposes of managing and accessing born-digital records. This could have particular application for responding to Inquiries where there is often increased public scrutiny and a need to justify decisions and responses.

5.4. Lesson learned 4: Identifying personal information

What is the challenge?

Born-digital records will often contain a range of sensitive information that should be redacted or closed from release to The National Archives. These exemptions will be made in accordance with the Freedom of Information Act and Data Protection Act. Such information includes, for example, the names and contact details of individuals and personally-sensitive financial details. Given that around 75% of exemptions to release relate to personal information, this is an area of priority and an opportunity to solve a significant challenge.

The challenge is not limited to digital transfer to The National Archives. The open data and transparency agendas require the need to screen information for personal sensitivities. Information security concerns mean that organisations need to understand the information they hold. Furthermore, the response to and release of information from Inquiries provides additional challenges around born-digital records.

What we learned

If you know what you are looking for and how to characterise it, the tools are powerful search engines that can find, highlight and (in some cases) automatically redact

- documents that contain specific keywords or names against a white list of names
- regular expressions (e.g. email addresses, credit card numbers, UK NI numbers, UK mobile or fixed numbers)

[24] Peck, A., (2011) 'Search, Forward. Will manual document review and keyword searches be replaced by computer-assisted coding?', *Law Technology News,* https://openairblog.files.wordpress.com/2011/11/peck-search-forward.pdf.

- names, locations or organisations based on natural language processing (including semantic dictionaries, document layout)
- conceptually similar sensitive documents based on the outcomes of the clustering and/or categorisation functionalities of the tools.

While the above can be achieved, it is the result of an iterative process. The tools need to offer the end-user the ability to feedback, exclude false positives and include missed items, which will allow accurate results to be increased over time.[25] This iterative process underlines the benefit of using a tool and/or supplier that allows the end user to add known omissions and exclude false positives on their own without having to rely on the supplier. Such reliance could be costly in terms of time and money and these arrangements should be clarified with the supplier at an early stage before procurement.

5.4.1. Gateway 2 - Sensitivity Review

If tools can be taught to identify known personal information (i.e. regular expressions), there are clear applications for the sensitivity review of born-digital records. They can support reviewers in removing content that should not be released. However, at the present time the tools cannot account for multiple variations on the spelling of a name. For example, Moammar Gaddafi has been shown to be spelt in 112 different ways.[26] The results also revealed difficulties with the placement of some characters. For example, a tool might find John Smith and J Smith but not J. Smith or Smith, J. False positives also persisted, with examples such as 'Kew, Richmond' returned as a personal name. This underlines the need for the end user to be able to iteratively interact with the tool to improve the results (without overly relying on the supplier). It also highlights that, similar to manual review, technology-assisted review is never going to be 100% accurate – departments will need to define and accept their risk appetite when using technology-assisted review.

Accuracy, in terms of technology-assisted review, means the right balance between precision and recall has been struck. This is sometimes referred to as the F-measure. Precision, in this instance, means the fraction of retrieved material that is relevant and recall being the fraction of relevant material that is retrieved. In simple terms, a high precision means that an algorithm returned substantially more relevant than irrelevant results while high recall means that algorithm returned most of the relevant results.

5.4.2. Inquiries

For Inquiries, there are likely to be requirements to search for named individuals and then potentially redact those names from publicly-available material. eDiscovery software tools can support these requirements, albeit within some of the limitations around the need to iterate and exclude false positives. For Inquiries, the personal

[25] A High Court of Ireland trial cited 25-50 iterations were normally sufficient to build an accurate predictive coding model. See: High Court of Ireland, (2015) 'Irish Bank Resolution Corporation Ltd versus Quinn', IEHC 175, www.bailii.org/ie/cases/IEHC/2015/H175.html.
[26] Gibson, C., (2009) 'How Many Different Ways Can You Spell 'Gaddafi'?', *ABC News,* http://blogs.abcnews.com/theworldnewser/2009/09/how-many-different-ways-can-you-spell-gaddafi.html.

sensitivities would be similar to a transfer of historic records but influenced also by their contemporaneousness.

Inquiry-related searches would still require considerable human input to achieve the maximum benefit from applying software tools. This includes utilising relevant keyword searches, predictive coding and following a documented process. However, it presents opportunities to utilise the human interaction at the most efficient and critical parts of the process and use some aspects of automation to reduce the overall burden.

5.5. Lesson learned 5: Procurement

What is the challenge?

Choosing the right software to support appraisal, selection and sensitivity review of born-digital records is dependent on procurement and other restrictions. This includes knowing the best products and suppliers to use. This will involve considerations around the security of personnel and data if needing to transfer born-digital records off site or give access to suppliers.

What we learned

We found a mature eDiscovery market with some well-established products clearly defined as applications to support legal litigation. In the wider software market we also found emerging tools with applicability to eDiscovery and other core target markets (e.g. financial investigations and financial due diligence). In this wider market we saw a lot of potential but less-developed solutions, or ones requiring additional development. In such cases the suppliers may have

- limited experience with working with government departments, which means they may have a limited understanding of the requirements and specificities of government departments – this can have important time implications
- additional flexibility and a willingness to make changes to adapt to new requirements, but are still in the process of developing their tools, even though they might 'sell' that they can deliver against specifications – this can have both cost and time implications.

5.5.1. Understand the business model

Given these differences between suppliers, markets and tools, it is essential to understand the supplier's 'business model' and their requirements before signing the contract. If a supplier makes a profit from selling the license for the product or provides an 'out–of-the box' product, you may have more opportunity to be able to use the tool without supplier involvement. This can be helpful when wanting to use the tool iteratively and if you are dealing with particularly sensitive information. In contrast, some business models are based on the development of bespoke solutions as well as the delivery of training for that solution, which could involve more cost.

Some suppliers need to take some or all of your data off site to 'train' the algorithm. This creates issues in terms of data handling, security clearances, and security of

the data and can generate costly delays. None of the above should preclude the adoption of such solutions and suppliers but should encourage organisations to be clear about both their and their supplier's requirements before committing.

5.6. Lesson learned 6: User interface

What is the challenge?

Knowledge and information management teams have extensive experience of handling paper records but born-digital records present new challenges. Therefore, the capability and skills within the teams may need to be improved.

In addition, the appraisal, selection and sensitivity review process will need to be iterative and repeatable. Business-as-usual transfers of born-digital records from departments to The National Archives will happen regularly. Inquiries will place *ad hoc* demands on search and retrieval capability, which will necessitate regular reviews of born-digital holdings.

What we learned

We learned that the user interface is as important as the quality of the algorithm. The software tools may be very powerful but unless they are accessible to the non-expert end user they will not deliver the benefits required. Extensive training and technical expertise needed to operate software tools will increase overheads and reduce enthusiasm for adoption. It is important, therefore, that before procuring any solution organisations should have a clear idea of their in-house capabilities and the skills required for using eDiscovery software.

5.6.1. Training the software

Software tools will need to be trained and kept up-to-date, the results of reviews regularly being scrutinised to ensure they fall within the departments' accuracy acceptance level. Sensitivities may change over time and tools will need to be retrained in light of these changes. To minimise costs after the initial setup the end user should be able to train the tools and reuse them without external support from suppliers.

5.6.2. Functionality to support Inquiries

Processes should be repeatable and auditable. For Inquiries, in particular, the process by which responses were arrived at (the methodology) may need to be repeatable and demonstrable to others to satisfy any audit or subsequent complaints. Accordingly it may be a requirement to be able to 'save' searches and queries run using the tools and apply these again. This could be to prove a previous outcome or to apply the same set of search criteria to a different collection of born-digital records.

5.6.3. Workflow

To support the user, a tool that fits the business workflow of appraisal, selection and sensitivity review is crucial. The ability to process and save workflows is likely to be important. Therefore tools should support logical business workflow processes that

reflect the way users want to work and can be recreated easily. This will enable effective training guidance to be produced and enhance auditing capability.

5.6.4. Agreeing wording

Understanding the wording and terminology used by software providers to describe the capabilities of their tools is important, as suppliers use different words for the same functionality. Without understanding the user experience, the underlying technology will be irrelevant so this requirement should be given appropriate consideration.

5.6.5. Test drive before procuring

Receiving a demonstration and then a trial to road test the software tools will be invaluable for testing the user interface and system functionalities. This should form part of the pre-procurement processes to ensure that end users can quickly pick up the tool. At this stage all key functionality should be in place so that it can be tested.

5.7. Lesson learned 7: Collaboration with other teams

What is the challenge?

While appraisal, selection and sensitivity review is a largely information management-driven process (albeit requiring consultation within the organisation and in some cases beyond), born-digital records require the input and skills of other business units. Primarily, this includes the information technology teams and potentially procurement experts.

What we learned

It is important to work with information technology colleagues to ensure the digital infrastructure can support the deployment of the software. There will be a number of considerations, including information security and whether the records will need to be migrated to an appropriate platform to enable interrogation. If the chosen product and supplier utilise cloud storage, the security and migration considerations will be important. There may be greater time and resource implications if the departmental information technology platform is outsourced to third-party suppliers.[27] Information technology and security colleagues should therefore be involved at an early stage in the planning and procurement process.

5.7.1. Information technology and procurement team involvement

eDiscovery software, as with other software products and services, can be procured in a number of ways. Information technology colleagues should be utilised to provide expert input to decisions and procurement professionals will be needed to examine the options. This will be important in procuring a cost-effective solution. Factors to consider are whether the software tools require the support of a third-party vendor to manage or can be bought directly from the product supplier. A decision will also be required around whether to buy a solution that can be accessed all the time or only

[27] Research by The National Archives shows that, out of 19 of the key government departments, 76% outsourced their information technology. The National Archives (2016) 'The Digital Landscape in Government 2014-2015. Business Intelligence Review', p.31, http://www.nationalarchives.gov.uk/information-management/manage-information/our-research/.

when needed. Further options include whether the software is deployed on the premises of the organisation or are cloud-based with storage provided off site.

5.7.2. Check for existing eDiscovery users

There is an opportunity to leverage existing expertise and cost efficiencies if eDiscovery tools are already deployed in the organisation. Legal or other teams, including data analysts, may already use software. This can help to overcome information technology challenges around security and deployment. Business cases can also be enhanced by demonstrating multiple purposes for expenditure of eDiscovery software. Therefore, it is important to check whether the organisation is already using eDiscovery licences for another purpose.

5.8. Lesson learned 8: Confidence in technology-assisted review

What is the challenge?

The challenge of born-digital records is new. Paper-based review processes are well-established and there is concern about meeting the digital challenge. These concerns include giving technology a greater role in the appraisal, selection and sensitivity review processes. In order to overcome the concerns it is likely, in the short term at least, that technology-assisted review will need to be validated against paper processes. This will help to build confidence in the use of technology to support human review.

What we learned

There have been some concerns over the reliability of the tools compared to a manual process, especially with regards to sensitivity review. There has always been an assumption that human file-by-file review is the gold standard when it comes to mitigating risks involved with sensitivity review – but this may not be accurate. The software trials revealed that the reliability of some of the software tools tested was inconsistent; however this does not prove that manual processes are superior.

5.8.1. Technology-assisted review can be more accurate than manual review

As opposed to a large number of human variables, technology has the potential to deliver more consistent results. For example, where humans tend to scan documents, meaning there is a risk of missing information, technology reads every word. This has led the 2007 Sedona Conference on the advancement of law in the United States to conclude that:

> [T]here appears to be a myth that manual review by humans of large amounts of information is as accurate and complete as possible – perhaps even perfect – and constitutes the gold standard by which all searches should be measured. Even assuming that the profession had the time and resources to continue to conduct manual review of massive sets of electronic data sets (which it does not), the relative efficacy of that approach versus utilizing newly

developed automated methods of review remains very much open to debate.[28]

Subsequent research into the comparison between human review and technology-assisted review has concluded that the latter can be better.[29] This has led to work in the Legal Track of the Text Retrieval Conference (TREC) on information retrieval methods to support legal eDiscovery.[30] Such research has increasingly led to a presumption in the United States legal process that, where appropriate, technology-assisted review should be used to reduce review costs and to cope with high volume.

5.8.2. The eDiscovery process is accepted in the legal field

The eDiscovery process is accepted in the legal field and technology-assisted review is becoming more commonplace within that process. It has been extensively researched and commented on by professionals and academics. It has withstood scrutiny in court. With this in mind, a system that meets these standards should also be deemed acceptable for the purposes of appraisal, selection and sensitivity review. Iterations to improve the accuracy of eDiscovery tools and positive demonstrations of accuracy when compared to manually-reviewed subsets of born-digital records will breed confidence in the use of software tools to support human reviewers.

6. Conclusion

This report has summarised the key lessons learned from testing and research around eDiscovery software and technology-assisted review. It concludes that technology-assisted review using eDiscovery software can support government departments during appraisal, selection and sensitivity review as part of a born-digital records transfer to The National Archives. This support also extends to responding to Inquiries and other activity such as replying to Freedom of Information Requests and information management, digital continuity and information security more broadly. We summarised these findings into eight lessons learned:

1. Understanding born-digital collections at a high level
2. Reducing the amount of information to review
3. Extracting meaning
4. Identifying personal information
5. Procurement

[28] Quoted in Grossman, M., and Cormack, G., (2011) 'Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review', *Richmond Journal of Law and Technology,* Volume 17, Issue 3, p.3.

[29] Grossman, M., and Cormack, G., (2011) 'Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review', *Richmond Journal of Law and Technology,* Volume 17, Issue 3.
Oard, D., Baron, J., Hedlin, B., Lewis, D., and Tomlinson, S., (2010) 'Evaluation of information retrieval for E-discovery', Artificial *Intelligence and Law*, Volume 18, Issue 4, p.7, http://terpconnect.umd.edu/~oard/pdf/jail10.pdf.
Peck, A., (2011) 'Search, Forward. Will manual document review and keyword searches be replaced by computer-assisted coding?', *Law Technology News,* https://openairblog.files.wordpress.com/2011/11/peck-search-forward.pdf.

[30] For information on the TREC Legal Track see: http://trec-legal.umiacs.umd.edu/#about. For further information also see 'The Decade of Discovery' (2014) documentary by 10th Mountain Films at https://vimeo.com/ondemand/thedecadeofdiscovery/107485099.

6. User interface
7. Collaboration with other teams
8. Confidence in technology-assisted review

eDiscovery tools and processes are not a silver bullet that will provide an immediate out-of-the-box solution. No one product or provider can offer a fully automated process with a guarantee of 100% accuracy. Risk appetites will differ between government departments on what constitutes appropriate acceptance criteria. However, based on an iterative process that requires end-users to engage with the technology, it offers ways to prioritise and reduce the volume of digital records that will have to be manually reviewed. This will support work for the digital transfer gateways, Freedom of Information responses and replying to Inquiries.

The eDiscovery software market is mature with some well-established products, but in the wider marketplace we also found emerging tools, often with potential but less-developed solutions. Suppliers also had differentiated business models, which would have different implications for departments in terms of cost, time and data handling implications.

The acceptance of technology-assisted review by the legal profession in the United States and more recently in Ireland is growing. This offers a level of reassurance in the reliability, accuracy and consistency of utilising technology to support review processes.

In the eDiscovery model we have a logical process that The National Archives will be amending for use by departments to support appraisal, selection and sensitivity review. Such a process will help government departments plan their born-digital records processes and provide a robust audit. This will support digital transfer and responses to Inquiries as well as information management and information security more broadly.

In general, manual review approaches utilised for paper records do not translate well into a digital environment, given the volume and lack of structure of born-digital records. Born-digital records also offer opportunities to utilise technology to meet the challenges in a way that was not possible with paper records. Research and the experience of the legal profession have demonstrated that technology-assisted review is both possible and can be better than human-only review. Future appraisal, selection and sensitivity review of digital records will require the assistance of technology and, among the current options available on the market, eDiscovery tools are the most mature and share comparable requirements.

7. Next steps

The next steps for The National Archives will involve refining the understanding that the eDiscovery software trials and associated research have produced. The National Archives will:

- Produce guidance for Gateway 1 (Appraisal and Selection) and 2 (Sensitivity Review) of the digital transfer process.

- Engage with stakeholders to produce an appraisal, selection and sensitivity review process based on the eDiscovery model that can be adopted by government departments. This will include understanding the acceptable levels of risk associated with using technology to support review processes.
- Develop outcome-based functional requirements (user stories) to assist in the identification and procurement of eDiscovery or similar tools.
- Continue to collaborate with academics and suppliers to further understand the potential application of software tools. This includes identifying the specific algorithms to best support particular types of born-digital record collections and make these findings available to government departments.

This will be an iterative journey that may need to be refined as processes are trialled with government departments and issues such as volume increase. We also expect technology and academic research to progress, and confidence in using technology-assisted review techniques in government to increase.

We will continue to work with the Cabinet Office and the Government Digital Service. In addition, we will continue collaboration with other centres of expertise within government and beyond to enhance methods and tools. In this respect we would like to encourage participation in our research efforts from other stakeholders. This could include piloting processes, providing sensitivity-reviewed sample data sets and commenting on guidance and requirements. By continuing to explore and test the possibilities, we will produce a range of support for government to meet the challenges of born-digital records.