

## From Start to Finish - what actually happens to my clients' data when I instruct a software and services provider?

By Chris Dale of the e-Disclosure Information Project and Deborah Blaxell of Epiq Systems

### The purpose of this paper

The purpose of this paper is to help lawyers and their clients understand what actually happens when they instruct a litigation services and software provider. Marketing materials inevitably reduce the processes to headings and bullet-points using a conventional and limited vocabulary; technical specifications are just that. Neither leaves a potential buyer of services with a clear understanding of what happens once instructions are given to collect data for litigation, for responding to a regulator's inquiry, or for internal investigations. What human and computer activities are embraced by the word "process"? How does the data get from the client's systems to the screen of the reviewing lawyer? How does it all relate to the client's objective?

Experienced users of such services know all about this, and make their choice of provider by reference to quality of service, proven reliability and price. Those facing their first case involving electronic documents may struggle to understand the concepts. These are best explained by simple descriptions of what actually happens from the moment instructions are given.

### Epiq and the E-Disclosure Information Project

This paper is written by former commercial litigation solicitor Chris Dale of the UK-based e-Disclosure Information Project<sup>1</sup> and by Deborah Blaxell, Legal Consultant at Epiq Systems<sup>2</sup>.

Epiq Systems is a consulting and software company specialising in the management of electronic documents and data for litigation and for regulatory and internal investigations. It has its own processing and document review applications, eDataMatrix and DocuMatrix, but will choose from a broad range of processing and review applications according to the requirements of the client and the case. Epiq offers a full range of services – consultancy, forensic collection, analysis and prioritization, and managed review.

Epiq's processes and methodologies are derived from years of R&D, from the experience of legal and litigation support staff drawn from Magic Circle law firms and major corporations, and from the input of client advisory boards. It is committed to the education of judges, companies and lawyers in the use of technology and techniques available to assist with the efficient management of e-disclosure exercises.

The eDisclosure Information Project brings objective and informed comment to lawyers, judges, suppliers and clients aimed at encouraging the better use of technology in eDisclosure. The educational objectives of the Project are sponsored by many of the world's leading providers of software and services, including Epiq.

### Executive summary

This paper takes a typical scenario – employees suspected of wrongdoing - and describes what Epiq and other top-tier providers of technology and disclosure management solutions (a "Provider") would be expected to do in response, from immediate discussions to identifying the presumed scope of the investigation and making a plan, through preservation and collection and into the various stages that follow. The steps known as processing, analysis and prioritisation tend to blur into one, and it is easy to lose sight of the client value that each of these brings before the review stage. The paper goes on to consider alternative ways of conducting the review when, as in this scenario, the opportunity to allocate resources may be denied by the urgency.

Whilst this paper is written from a UK perspective, many of the principles apply in any jurisdiction.

### Scenario

---

<sup>1</sup> <http://www.edisclosureinformation.co.uk/edisclosureproject.htm>

<sup>2</sup> <http://www.epiqsystems.co.uk/>

A corporate client suspects that a number of its employees may be involved in wrongdoing and wishes to conduct an internal investigation to assess whether it needs to self-report to a regulatory authority. The company has identified four individuals who may be involved in the wrongdoing and instructs a Provider to assist with the collection of their data.

This scenario differs from litigation in a number of ways, not least because of the extreme urgency which an investigation of this nature often imposes on businesses. One characteristic is that the company is not merely complying with a discovery request (in the US) or a defined disclosure obligation (in the UK), but needs to know, and as quickly as possible, how big the problem is – how many people are involved, how long they have been acting in this way and what the implications are for the business. Another complication, covered in the paper, is the possibility that the exercise may have to be kept from those whose help would otherwise be called upon.

#### **Taking instructions:**

Once the corporate client or its lawyers have contacted a Provider, the Provider's team has a meeting or conference call with the client to understand the nature of the problem. The Provider will need to understand the time-frames involved, the objectives of the project, the risks involved, the client's budget and the resources available to assist with the project. If the project is large or complex enough to warrant it, the Provider will map out a solution blueprint – a document proposing a strategy for undertaking the management of the data from identification through to analysis and review. It is very important at this stage that the Provider, the client and the client's lawyers communicate fully to ensure that each party understands the roles and objectives of the other parties. Where the Provider has been fully briefed and understands the overarching requirements of the client and its lawyers (insofar as is possible), it is more likely that effective results will be achieved. The Provider will also need to understand, for example, whether the investigation must be covert or overt – that is, whether the target individuals are to be made aware of the investigation and whether the IT department can be involved in the collection. Factors include the risk of data destruction and the risk that employees will be tipped off.

It is important for the Provider to gain an understanding of the data sources and the relevant keepers of the data ("the custodians") as soon as possible. The Provider will also need to understand the time-frames for which the custodians' data is relevant and any particular time-frames to which the client or its lawyers are working.

The obvious starting points in identifying data sources will include the following:

1. Email from email server
2. File shares
3. Work PCs
4. Personal PCs, mobile phones and PDAs
5. Data from company mobile
6. Removable media
7. Hard copy (notebooks, diaries, wall calendars, etc.)
8. Tape

The range of potential data sources may also include, for example, instant messaging and various forms of social media. Those giving the instructions will not necessarily be aware of all of the means by which their employees communicate.

Who are the people that matter, apart from the four suspects? An iterative mixture of discussion with the client and data searching is required, as human input suggests avenues for searching which may in turn prompt questions. For example, sampling a particular custodian's data may show that he or she is irrelevant (in which case the search can be narrowed to exclude that individual). Alternatively, an initial search may show that the custodian has been communicating with other individuals whose data must also be searched (thereby extending the scope of the search). Such decisions may need client input – a simple question may justify or dismiss an apparently promising line of enquiry.

#### **Preservation:**

There is no second chance to preserve data and a prudent Provider will usually advise the client and its lawyers that it is important to preserve widely – preserving only a narrow amount of data early in the investigation exposes the client to a potential risk that crucial evidence relating to issues yet to be considered may be permanently lost.

Once the data is preserved, the client and its lawyers will be free to dip into the ring-fenced data sources as required, safe in the knowledge that the data has been secured.

In this regard, a Provider will advise upon the kinds of data to consider, based upon the information provided to it as part of a consultative process with the client and its law firm. The Provider is likely to:

- Identify the key individuals involved, casting the net broadly in the first instance;
- Establish whether there is a standard IT deletion policy in place and whether it can be suspended for particular individuals;
- If necessary, take a forensic copy of these key individuals' data sets (leaving the original data unmodified);
- Conduct some analysis of the data at the earliest opportunity, to confirm that the preservation exercise has been carried out broadly enough;

Identify back-up tapes for the relevant period and secure them (i.e. take them out of back-up rotation) to ensure that the information stored on the tapes is secured.

The benefits of this approach are obvious. The preservation of evidence is key to any investigative approach. By preserving broadly and early, the client and its law firm will be better able to make informed decisions about the nature of the issues involved in an investigation as they evolve. Early and substantive preservation protects against data / evidence loss. Clients are often concerned that a wide preservation exercise will add unduly to the costs and time spent on the disclosure exercise. Generally, however, only relatively modest costs are involved in the actual preservation exercise. The amount of data that will actually be processed and reviewed is likely to be much smaller than that which is preserved.

#### **Collection:**

*Once identified and preserved, the data will need to be collected.*

Ideally, this is a joint exercise between the Provider and the client's IT team. Different considerations apply in particular circumstances – at one extreme, security concerns (e.g. where government defence contracts are involved or where particular government security clearances are required) may bar hands-on involvement by Providers that do not have the requisite clearances; at the other extreme, the IT department may be connected with the people being investigated. The client may have a bespoke information management system (e.g. SAP) in place; it might have anticipated the requirement for forensic collection or might have no systems or teams at all.

These factors will dictate whether a Provider attends at the premises or has a merely advisory role. Either way, the Forensic Consulting team will advise upon the best method of preservation and collection in the circumstances, set a timetable for the collection exercise and ensure that records are kept, evidencing chain of custody.

During a collection exercise, disruption of normal business processes can occur, especially if data needs to be collected from live servers. A Provider's typical approach should be to image this data and filter it off-line which minimises disruption and provides a comprehensive data store, reducing the need to go back on site should the terms of investigation shift.

The question who collects the data will be tied to whether the collection is being undertaken covertly or overtly. In an ideal situation the IT team, advanced users of proprietary systems (for example a financial database), or the custodian should be involved in any collection. They have intimate knowledge of data locations and access permissions and will be able to suggest data sources which may not instantly occur to the investigation team.

Ideally, the IT team will assist with the harvesting of data such as email (unfiltered) into archive file formats (PST, NSF or equivalent), access to file server locations, or the bypassing of encrypted files which may be used on custodian laptops or BlackBerries.

Advanced users of proprietary systems may assist in running searches within such systems and creating reports that can be exported for analysis by the investigation team at a later time.

Where possible, full disk images of custodian PCs, laptops, or captures of data from BlackBerries and similar devices will be carried out by the Provider or an independent consultant. However if the client has an in-house capability, the activity may be carried out by them in consultation with an independent consultant.

Hard copy documents should be scanned or in rare circumstances, photographed, either by external consultants or in house, with the guidance of an independent consultant. It is important to translate the physical relationships of paper (e.g. a paperclip) into parent and child relationships on a database. The quality of scanning will have an impact on OCR (optical character recognition) quality, and hence on the effectiveness of analysis and search terms thereafter.

If a covert collection is needed the collection schedule is based on the available windows of opportunity i.e. periods when the PC can be accessed without the employee's knowledge. After the covert aspect of the exercise has been undertaken, the task of imaging and examining the data is no different to the standard overt methodology.

A consideration when dealing with covert collections is that the subject or IT team may need to be given a diversionary assignment to force a window of opportunity to become available.

The benefit of a covert collection and examination is that it gives the client time to test their concerns and to prepare for any media or client fallout that bringing the investigation out into the open may cause.

#### Processing and Hosting:

Once the data has been collected, it must be processed. Processing is the stage which is least understood by those new to electronic disclosure, perhaps because it is a generic term with wide connotations. In the present context, it involves a set of computer applications which extract the maximum information from the data which has been collected and then uses that information to guide decision-making as to the next steps. There is a wide range of processing technology available.

#### Deduplication:

A top-tier proprietary processing engine will extract metadata, text and embedded objects from the data and discard files that are obviously unneeded, such as system files and exact duplicates. The removal of duplicates is a means of reducing the document population electronically. Duplicate is a broad term, but when processing there are a number of key approaches. The first is to de-duplicate exactly, on the basis of a file algorithm or hash value. The advantage of doing this on a sophisticated database system is that you can keep placeholders to show the provenance of an email (e.g. that it belonged to 5 different custodians).

When working this way, there are two main options – to de-duplicate *within* custodian (an effective method if dealing with multiple versions of the same data) or *across* custodians (minimizing the number of documents to review, but clearly having major implications if conducting a custodian-by-custodian document review). Exact duplicates can be too exact, however – the same email in Lotus Notes and Outlook format will not be the same electronically. To overcome this problem, the Provider may, for example, use email threading, which can identify duplicate emails on the basis of the text itself. It can handle foreign languages (the technical term is Unicode compliant – so will include Russian and Chinese) and a wide range of data formats. The end result is a much smaller dataset and a series of reports which give an informed assessment of the case.

#### On-site and off-site solutions

Processing is usually undertaken by the Provider's production teams at their premises. However, if there are jurisdictional challenges, data privacy issues or if the client is particularly sensitive to confidential data leaving its premises, the Provider may be able to provide a solution to process data on the client's site, using a combination of third party software and proprietary software.

Hard-copy documents may be scanned into the system to enable them to be searched alongside electronic documents. This can also be done on- or off-site.

#### A staged approach:

Whilst it is obviously good to be able to process all the data at once, urgency may dictate an immediate start with the first (the most obvious, most important or most easily collectable) sources, adding others as they come in. Urgency is not the only reason for a staged approach. It makes economic, as well as practical, sense to begin with the data which is most likely to contain the evidence, moving more widely only if necessary. As the first selected data goes through the early stages of processing, it may become clear that other custodians, other data sources, or a wider date range is suggested by something ignored on the first pass. It may transpire, for example, that a suspect has been in correspondence with someone else whose involvement cannot be explained by normal working patterns. This approach is becoming acceptable even in civil proceedings; whilst the rules may appear to require full disclosure of everything potentially relevant, the informed use of judicial discretion may point to a narrower ambit to begin with. This

approach was taken by Senior Master Whitaker in his judgment in *Goodale v Ministry of Justice*<sup>3</sup> where he identified four custodians to be the subject of the initial disclosure, making the express presumption that all potentially relevant data had been preserved.

Once processed, the remaining data either moves to a hosted platform or is placed in a load file and exported to an alternative system, for example, at the law firm, that enables the review team to look at the documents.

### Analysis

Analysis is another seemingly technical term which embraces a range of easily understood concepts. The recurring theme of value to the company or its lawyers is that documents of a like kind can be grouped together:

- *Clustering* identifies documents with shared terms or words and groups them together;
- *Categorisation* identifies documents which are conceptually related to each other, grouping them in category folders and enabling an immediate answer to the question "show me more like this";
- *E-mail threading* takes all of the messages from a single e-mail conversation thread, together with their attachments, and allows the user to see the "inclusive" e-mail which contains the rest;
- *Near-duplicate identification* collects documents which are nearly the same as others so that they can be reviewed together.

These are all features which can make the ultimate review experience a more efficient and cost effective exercise. At best they allow the removal of whole categories from consideration. At the least, the fact that the same reviewer sees all related documents at the same time prevents inconsistency and duplication of effort.

This stage can be done on- or off-site. Providers like Epiq, with strong project management and consulting teams, are on hand to advise upon the most appropriate accelerative features for use in each disclosure exercise. Systems with this functionality enable the allocation of the data to individuals within teams of reviewers, thereby assisting those organizing the review to regulate the flow of documents to the appropriate reviewers, matching skills and seniority to difficulty and importance, and measuring progress.

### Forensic (subject matter) examination

Forensic examination is the process of presenting a number of assertions or questions about a Custodian, their alleged actions and the data or technology at their disposal. Epiq has a team of Forensic Consultants who are able to examine the collected material with a view to establishing whether allegations are supported by evidence, and whether the questions can be answered (either positively or negatively), then report on their findings.

In our example it may be necessary to consider whether web-based email was used by the subjects to communicate. Key questions that a Forensic Consultant will consider include whether any of the conversations can be recovered and whether any company material that was disseminated via this means was subsequently used by the suspects.

Whether a detailed examination is required will depend upon the circumstances. Forensic examination and analysis can happen either before or at the same time as the document review, or at any stage during the process when an issue (question) arises for which this type of detailed analysis is appropriate.

### Prioritisation

Epiq's project management and consulting team can also take lawyer input into a sub-set and apply the results across the rest (Epiq calls this "IQ Review"; some other Providers have similar functionality). The aim here is not merely to decide whether documents are relevant or not, but to rank them in order of priority of the most likely to be relevant documents.

Put briefly, a lawyer or subject-matter expert who is familiar with the issues of the case is given small batches of documents from the culled-down set which has resulted from the earlier stages. The expert then grades each document responsive or non-responsive according to its relevance to the issues. The decisions made about these documents are fed back into the system which, after a number of iterations depending on the size and complexity of the dataset,

---

<sup>3</sup> <http://www.bailii.org/ew/cases/EWHC/QB/2009/B41.html>

applies those decisions across the rest. A relevance threshold can be set, and the results can be checked by sampling those which fall either side of the threshold.

Multiple benefits follow from this approach. Quite apart from the relegation of documents which have no or a very low relevance score, the ability to feed the most relevant documents to the appropriately-skilled persons right at the beginning of the review has obvious benefits. This kind of investigation depends on finding quickly the handful of critical documents which demonstrate unlawful activity or even, perhaps, the so-called "smoking gun" which puts the case beyond doubt.

### The result of the pre-review stages

If we revert to our scenario, in which a client needs urgently to know about the activities of a group of its employees, processing and analysis are stages which amply repay the time and cost spent on them where the aim is to get the eyes of decision-makers on to the smallest set of relevant documents in the minimum time in a form which encourages speedy review. Prioritisation further reduces the set to be reviewed, and puts the result into an order which reflects the degree of relevance to the problem to be solved.

Some of these stages require little human input – identifying duplicates and discarding system files is mostly a computer function. Others require choices – where to set thresholds for identifying near-duplicates, for example. Yet others involve equal dependence on human and computer functions, such as prioritisation, where skilled people set the parameters and the computer applies them. All of the stages benefit from the iterative involvement of clients, lawyers, and the Provider.

### Review

In broad terms, there are two different approaches to document review. Deciding what is appropriate involves a number of factors including:

- The number of documents to be reviewed
- The nature of the investigation or litigation
- The timeframes imposed by the court, the regulators or (in the example we are working through here) the perceived implications of the wrongful activity
- Cost constraints

Document review is inevitably the most expensive stage in a typical exercise because it traditionally involves the reading (often at lawyers' hourly rates) of all documents which have survived the prior processes described above. Lord Justice Jackson drew attention to the expense of this stage as did Master Whitaker in *Goodale v Ministry of Justice* when he said:

*This is a prime candidate for the application of software that providers now have, which can de-duplicate that material and render it down to a more sensible size and search it by computer to produce a manageable corpus for human review – which is of course the most expensive part of the exercise. Indeed, when it comes to review, I am aware of software that will effectively score each document as to its likely relevance and which will enable a prioritisation of categories within the entire document set.*

The possible approaches are:

#### Review by in-house or external lawyers

The conventional approach, particularly in the context of formal proceedings such as litigation, is that the clients, or more usually their lawyers, set a team of paralegals and trainees to conduct a "first -pass" review. More substantive work, such as privilege review or issue coding, is undertaken by more experienced, and hence, costly associates. That review may be linear (that is, by moving from document to document in date or some other convenient order) or may take advantage of the grouping tools described above (e.g. bundling together documents that are connected by content, thread or degree of likeness, or that have been ranked for priority). Review tools include other review accelerators such as the ability to mark documents in bulk because of some common characteristic. They may also include the project management tools referred to above.

#### Managed document review

Not every company or law firm has the resources or experience to undertake a large or complex review at short notice. The lawyers, or their clients, may instead choose to outsource the review, that is, to engage the services of a process-driven document review company. Off-shore review allows database web-access to foreign-qualified lawyers around the world. Access to overseas lawyers in lower-cost economies mitigates the often very high cost of the traditional charging model in the information society, yet raises a host of other issues. Differing time zones, data protection issues, language and cultural differences, lack of nuanced understanding of legal concepts such as privilege, loss of control, and quality assurance are just some of the concerns raised by this method of review.

As an alternative, Epiq and a small number of other Providers offer an on-shore document review service bringing together the best of both worlds and overcoming the weaknesses in each as well. Epiq's service provides qualified lawyers with relevant subject-matter and document review experience to work with the law firm and its clients to undertake all or the first pass of the review. This can be done on-site at the law firm or the client's premises. Epiq also has bespoke document review facilities offering secure, transparent, measurable and defensible managed review services.

### The Output from the Exercise

The nature of such a project means that the client and its lawyers can be constantly involved in the interim stages and kept up to date as matters develop. At the end, it needs reports both to make a decision and, if necessary, to underpin that decision with evidence gleaned from the exercise. Epiq provides a range of detailed reports outlining the various methodologies used and recommendations and findings made during the project

### Lessons learnt

Whatever the purpose of an exercise of the kind described above, much of its value lies in the lessons which can be learnt from it. One of the supplementary benefits which clients should require is a report on the exercise and recommendations based on what happened. It may be, for example, that the collection was hampered in some way, either because of the client's technology or because of its human processes which had not hitherto been put to the test. Part of the value of this lies in ensuring that the same problems do not arise next time; even more value lies in being able to head off future problems of this type.

### Summary

There is often little time to make decisions when an issue arises of the kind described in the scenario above. The suspicious conduct may affect the company's cash or other physical assets; it may involve a potential self-reporting obligation whose timing is critical; it may have wide repercussions vis-a-vis a regulator, a prosecutor (e.g. under the Bribery Act) or because of stock exchange implications. The overall task is broadly the same as for litigation, but the urgency and the range of implications is often much greater. In our scenario, one of the requirements was the urgent preservation, identification and management of the company's information sources. This involved, in varying degrees, the application of in-house resources and third-party Providers of software and services. The starting point for the business was understanding what the exercise involved and knowing who could provide the relevant services so that the business was able to conduct its investigation in the minimum time.

More generally, an effective response to urgent regulatory investigations, internal investigations, litigation and the like requires preparation. There are two elements to this preparation. One is risk analysis, that is, the identification of potential strengths or weaknesses of the company's position, likelihood of any negative issues arising, and their implications for the business. The other is planning the reaction to any such threats, defining and establishing responsibilities, and identifying the internal and external resources which can be applied to resolve them. The tools and techniques described above, whether from Epiq or other leading Providers, are an important part of the establishment of an effective response. However, it is the implementation of a strong process-driven approach, devised with the assistance of solution-driven experts that will ensure that, whatever the matter, the company's response is confident, timely, defensible and proportionate.