Equivio Zoom for Information Governance – Enabling Categorization to Reduce Risk and Cost and to Find and Extract Value from Corporate Data Stores

White Paper

by

Chris Dale of the e-Disclosure Information Project

equivio
zoom in. find out

This paper is written by Chris Dale of the UK-based eDisclosure Information Project in conjunction with Equivio. Equivio made its name developing software for eliminating data redundancy, primarily for eDiscovery purposes, that is, for the reactive task in which lawyers and investigators need to identify quickly those documents which matter for the purposes of litigation or an investigation.

Equivio has now applied that technology to the prior and wider problem of information governance, and specifically to the categorisation of documents as they sit on a company's systems, not least so that unwanted documents can be deleted in a defensible manner.

In an Information Governance context, “categorization” means categorization by retention categories. These retention categories are generally defined by the company, and there may be tens or even hundreds of such categories. Each category has a retention period attached to it, which may be for 1 year, or 2 or 5 or 10 or 25, depending on legal requirements and the company’s view.

If you can achieve this, while also identifying data that can be defensibly deleted, the eDiscovery problem diminishes in consequence. Equally importantly, the categorisation enables companies to find and extract value from their data stores.

There is more to this than simply applying an existing eDiscovery application to new uses - there are differences of scale and purpose which require adaptation of both approach and of technology.

No matter how good the technology used for electronic discovery, data volumes continue to grow, and the time has come to carry the battle against redundant and irrelevant data back into the corporation’s data stores.

WHAT IS INFORMATION GOVERNANCE AND WHO IS INVOLVED IN IT?

No matter how good the technology used for electronic discovery, data volumes continue to grow, and the time has come to carry the battle against redundant and irrelevant data back into the corporation’s data stores. Electronic discovery is a high-profile end-use of information, but it is not the only one. A company’s data is simultaneously an asset and a source of risk, and one in which a wide range of interests should be represented.

The Information Governance Initiative, of which Equivio is a member, lists 13 areas of a company’s business which are embraced by the term Information Governance. They are:

- Information security
- Data science
- Electronic discovery

- Business management
- Compliance
- Business intelligence
- Analytics
- Records management
- Finance
- Audit
- Privacy
- Risk Management
- IT and Infrastructure Management

What all these interest groups have in common is that growing volumes of data increase the time, cost and risk inherent in their work whilst reducing the benefit which might be derived from it. The risk/benefit balance is more complex than one might think: discovery is generally seen as a source of risk, but there are benefits in finding the critical documents quickly; business intelligence is largely a benefit, but the failure to find reusable content – a good PowerPoint or some relevant statistics – may lose a competitive advantage or the opportunity to make a timely point in negotiations or in a dispute.

Hitherto, the development of analytics tools like Equivio Zoom has been largely for reactive purposes – finding urgently the documents which respond to a discovery request or which are required by a regulator. The time has come for companies to think more positively and to apply this sophisticated technology to the task of reducing the volume of data proactively rather than merely because a subset of it is needed now.

That is what Equivio Zoom for Information Governance is designed to do.

DESCRIPTION OF THE PROBLEM

A company has too much data. For years it has not really seen this as a problem, buying new servers and storage devices as required and (perhaps) migrating material from old devices to new ones. Its lawyers have repeatedly advised the company to keep everything "just in case", and users are both reluctant to give up material which they might need one day and too busy to address a problem which has, in truth, passed beyond the stage where the occasional "deletion day" will solve it.

Hitherto, the company has considered only the bare storage costs in assessing the burden on the company. Now, however, it has begun to look more closely at the costs.

IT has added up the annual cost of maintenance charges, software costs, security measures and the not inconsiderable salaries of the people who manage it all. Legal is making ever more frequent demands in relation to compliance obligations of various kinds which are requiring the company to produce documents. The eDiscovery component of litigation is matching or exceeding the cost of legal advice.

The reported assets of the company include large amounts of valuable intellectual property, but no one knows what it comprises or how to find it. HR is making ever more frequent requests for information as it responds to demands from employees and external agencies. High-profile security breaches at other companies are causing concern as it becomes clear both that the company is holding much time-expired private information which is not required for the business and that there is no effective way to trap the unauthorised removal of data because no-one knows what ought to exist.

> Lurking unseen among the data there may be what might be called “gotcha” documents — documents which evidence some fact or conduct with potential to be embarrassing at least and incriminating at worst and which you were entitled to delete.

Lurking unseen among the data there may be what might be called “gotcha” documents – documents which evidence some fact or conduct with potential to be embarrassing at least and incriminating at worst and which you were entitled to delete. That opportunity passes when an investigation begins or litigation arises and the documents turn up in an eDiscovery collection.

On top of all this, business managers are complaining at the time it takes to find information which is critical to the company's ability to compete.

A decision is made both to reduce the historic volumes and to impose policies for the control of newly-created data. Both these tasks require some level of categorisation, both in respect of historic data and in relation to newly created documents. Someone actually has to do the work, and many attempts at volume reduction fall down at this point.

MANAGING THE PROBLEM

A document retention policy is a set of rules and procedures setting out how a company’s documents and data are to be stored and, in due course, when they are to be deleted. The policy should dictate where documents are to be stored, in what format, with what additional information, on what media and for what periods. Different types of documents may have different policies depending on their purpose, their originator and any statutory or other relevant elements. Provision should be made for backing them up, archiving them, and for ensuring both that they have a retention period and that they are disposed of when that period expires.

Companies have a range of choices in dealing with the categorisation of documents for retention. One approach, sometimes referred to as the “trusted custodian” method, requires employees to execute the policy by, for example, dragging and dropping emails and documents into relevant retention folders.

Such approaches may include a requirement to complete a form which indicates the document's categories and, perhaps, defines (or allows the user to define) a deletion date. This is all difficult, time-consuming and sporadic. Individual employees also tend to have their own interpretation of the retention policy. US Magistrate Judge Andrew Peck, a strong advocate of corporate information governance, observes that a salesman with a choice between making one more call to a prospect or completing some document categorisation forms is more likely to choose the former – and in many ways the company would rather he was making that call.

The opposite course (generally by default) is to ignore the problem, and simply retain everything more or less forever. This approach leaves the important documents buried amongst the rest. The most immediate time and cost implication for many US companies arises on discovery when all these documents must be identified and ploughed through for every case or relevant regulatory demand. As the preceding section showed, over-burdensome discovery may be the most obvious implication, but it is not the only one.

In between lies a middle position which is often worse than either of the extremes – companies define a policy which they are unable to execute. This is more than a mere management failure and a waste of resources. The deletion of documents pursuant to an appropriate policy is acceptable in most US courts provided firstly that none of the documents was (or should have been) subject to a legal hold and secondly that the policy is actually executed. The existence of documents which should have been deleted can seriously undermine a company's arguments when challenged before the court.

USING PREDICTIVE CODING FOR INFORMATION GOVERNANCE AND DEFENSIBLE DELETION

If predictive coding is an accepted way to weed out documents defensibly for discovery purposes, then why not use it for the same purpose in an information governance context? It is just as important to do the job defensibly, but there is, relatively speaking, leisure to develop processes and practices, not least of quality assurance, to ensure that the job is done properly, without having it done by lawyers at hourly rates and in a hurry. Furthermore, once the job is done, its results are permanent, whereas a discovery selection process may flush out the same irrelevant documents several times for succeeding cases.

In a discovery context, predictive coding aims to achieve two main results – the categorisation of documents as being responsive or not, and a prioritisation ranking which indicates the degree of relevance or of potential importance. The process begins with case strategists or subject-matter experts reviewing samples of documents selected by the system and coding them as relevant or not. The algorithm is trained by an iterative learning process using textual analysis to decide whether one document has characteristics shared with others. The results are then applied across the whole of the relevant corpus of documents resulting in the ranking categorisation referred to above. Sampling and other forms of quality assessment follow.

The use of predictive coding for information governance is more complex than this. One challenge is the multiplicity of categories – documents are not just categorised as “Yes" or “No" but must be put into one or more of multiple categories. If there are (say) 200 categories, then the IG predictive coding system must be able to train very efficiently in order to minimize human effort.

Secondly, the richness of documents varies. Richness, in this context, is the ratio of relevant documents to the total population. In a discovery context, the collection will usually have been refined at least by custodian, date range and file type, and the target – the question whether documents are relevant or not – is a wide one in the sense that every document will fall into one or other of those binary buckets.

For defensible deletion purposes, the outer scope may be unrestricted and there are multiple buckets; in addition, companies have many things which fit into no known category. A richness of 5% is acceptable for discovery; the average richness of document retention categories may be .05% or even smaller. The system has to take account of the percentage of useful documents out of the total population. You cannot just take predictive coding as developed for eDiscovery and use it for defensible deletion. The sampling techniques, statistical model and training strategies used for predictive coding in eDiscovery need to be modified to accommodate the IG environment.

WEIGHING BENEFIT AND RISK

The objective is to balance the risk of deleting material which should be kept against the risk and cost of keeping documents the company wants to delete. Different companies in different industries will take different views as to what the risk is – all companies are required to keep some documents, but the objective is to minimise the "dark side" of the data and maximise the retention of material which the company wants to keep.

There are human and strategic decisions to be made here, balancing a higher risk of deleting documents you should keep against the risk of keeping documents you want to delete. Companies have to decide where they want to be on a balance which sets degrees of risk against volumes deleted. Some companies may be happy to delete much more, but everyone is obliged to retain something. What mix of cost and risk is appropriate for this organization?

THE PROCESS

There are five stages to a records retention process using Equivio Zoom for Information Governance:

Step 1 Define the categories

The subject matter expert can use his or her prior knowledge of the categories and the kinds of documents that are responsive to those categories, to jump-start the full-scale training process with some seed documents.

The objective is to train the system to identify the category, and each category is set up with its appropriate retention bucket. Once you know the category to which a document belongs, you automatically know its retention bucket. There might be a one-year retention bucket, a five-year retention bucket and a bucket for junk which you can get rid of immediately. The categories, and the number of categories, will vary with the organisation and with the intended scope of the exercise.

Step 2 Train the system

Once the retention buckets are created, you can start training the system. Each training set consists of 40 documents, and users' coding decisions allow the system to recognize the characteristics of documents which fall into the pre-defined retention categories – the process is iterative and self-correcting, that is, the system will keep training until it knows what decision should be made about the target documents (the original use of the word "predictive" in this context is that the system can predict what decision a skilled user would make). The exercise continues until the system reports that it is "stable", that is, that it has enough information to apply the classification criteria across the wider set of documents.

Step 3 Perform QA

Once training is complete, the next stage is a guided QA process to verify and quantify the results. If errors are discovered, they can be fed back into the classifier, and further rounds of training ensue.

Step 4 The Decision Stage

The next step is batch calculation which calculates the category classification scores for each document in the repository. Once this is complete, the system allows the user to select the cut-off score for defensible disposition and deletion of documents. Each cut-off decision has different cost and risk implications. Using an intuitive visualization device, Zoom shows you the prospective result of taking different decisions so that the company can choose where it wants to be according to its cost and risk management profile. The system provides options for additional modes of QA to verify the user's decision.

Step 5 Apply retention rules

In this stage, the system invokes the cut-off point, as determined in the decision phase, to assign documents to retention categories, applies rules, as required, and sets the appropriate retention bucket for each document.

A PROOF OF CONCEPT

Law firm Drinker Biddle & Reath undertook a proof of concept (POC) exercise with Equivio's Zoom for Information Governance product. Their client was an international bank and the POC focused on a repository with 4 million documents. The exercise began with Drinker Biddle defining narrowly what types of documents were required for an "internal audit" category and a "regulatory requirements" category. In addition, a number of junk sub-set categories were defined. The seeding, training and QA phases took one day each, a total of three days for all the categories together.

> LAW FIRM DRINKER BIDDLE & REATH UNDERTOOK A PROOF OF CONCEPT WHICH SHOWED THAT OVER 45% OF THE DOCUMENTS COULD BE DEFENSIBLY DELETED, WITH A RISK FACTOR OF LESS THAN 5%.

Based on sampling of the data set, and projected against a full-scale project, the POC showed that over 45% of the documents could be defensibly deleted, with a risk factor of less than 5%.

CONCLUSION

Equivio's development of Zoom for Information Governance began with a predictive coding application which was already widely used and accepted by US courts for the task of deciding responsiveness for litigation under the US Federal Rules of Civil Procedure. It spent a year redeveloping the system for what is in many ways the more challenging task of deciding retention categories for information governance – more challenging not just because of the potentially larger volumes but because of the multiplicity of categories and the low richness of a typical corporate data set.

The resulting product has been the subject of two major proofs of concept, the one with Drinker Biddle described above and another with Epiq Systems. There are a number of other such exercises in hand or being planned, and this level of interest is itself evidence that companies and their lawyers recognize that the twin and closely-related problems of data volumes and dark data cannot be left unmanaged.

For most companies the problem has passed beyond the stage where individual end-user employees can be expected to manage historic volumes – in many cases the users will have long gone anyway, and such user-led approaches are fraught with subjective views, personal preferences and simple lack of time.

The technology of simple keywords is barely adequate for keeping track of a single user's output for a short period, as anyone knows who tries to keep an InBox in order. When the problem is multiplied across many users and long periods, something more sophisticated is required.

Equivio Zoom for Information Governance allows a balance to be struck between the duty to retain certain documents and the ambition to reduce overall volumes. The QA tools and processes which are an essential element of litigation discovery are equally important in the context of information governance.

Equivio reports considerable interest in the use of Zoom for Information Governance. This is unsurprising; the problem which it manages is one which weighs heavily on many companies.

Chris Dale
Oxford
August 2014

About Chris Dale of the eDisclosure Information Project

Chris Dale qualified as an English solicitor in 1980 after reading History at Oxford. He was a litigation partner in London and then a litigation software developer and litigation support consultant before turning to commentary on electronic disclosure / discovery. He runs the e-Disclosure Information Project which disseminates information about the court rules, the problems, and the technology to lawyers and their clients, to judges, and to suppliers. He was a member of Senior Master Whitaker's Working Party which drafted Practice Direction 31B and the Electronic Documents Questionnaire. He writes an authoritative and objective web site and blog on the subject and is a well-known speaker and commentator in the UK, the US and other common law jurisdictions.

http://chrisdale.wordpress.com
chrisdaleoxford@gmail.com

About Equivio

Equivio provides text analytics solutions for legal and compliance. Equivio offers Zoom, a court-approved machine learning platform for the legal arena. By enabling the defensible quantification of compliance decisions, Zoom has transformed the business process of e-discovery. Now Zoom is also leading the transformation of information governance. Zoom users include hundreds of leading corporations, law firms and government agencies. Zoom organizes collections of documents in meaningful ways, while quantifying and visualizing the decision space. So you can zoom out for the big picture. Or zoom in to find just what you need. Visit us at www.equivio.com or send an email to info@equivio.com.

Zoom in. Find out.