[ICN Training on Demand: Document Review]

[ICN TRAINING ON DEMAND: DOCUMENT REVIEW]

[Alex Schofield, Policy Adviser, UK Competition and Markets Authority]

ALEX SCHOFIELD: Hello, my name is Alex Schofield, and I'm a policy adviser at the United Kingdom's Competition and Markets Authority, the UKCMA. Welcome to this module which looks at document review. This module is part of the ICN's Training on Demand series on conducting investigations. This module will concentrate on planning and preparing for document reviews. We will be sharing approaches and experiences that we hope will give you some inspiration for dealing with the challenges that you face.

Our guides today will be Gary Bracken, Sophie Mitchell, and Chris Dodds from the Antitrust and Cartels Investigation Teams at the UKCMA. Their day-to-day work involves running antitrust and cartels investigations and they are experienced in running document review exercises.

We will also be hearing from colleagues at competition authorities from Brazil, Colombia, South Africa, Singapore, and the US, who will talk to us about some of their experiences and approaches in practice.

We're very grateful for their support and to all of the authorities that have helped with the development of this module. This has really been a team effort and we hope that the blend of experiences provided by different jurisdictions will give you, the audience, a lot of useful tips for document review.

In this video, we are going to talk about how to plan and conduct document reviews. What will we cover? Well, we're going to examine what you need to think about before the review, what to consider as you collate the

documents and prepare for the review, and how you might carry out the review.

With help from our contributors, we'll also discuss how competition authorities have managed the transition from a paper-based to an electronic case file and how document review platforms can help you, including how some of the newer technologies, such as predictive coding and machine learning, might assist in document review.

Finally, we'll consider some of the challenges that you might face in completing document reviews.

Before we get started, we wanted to make sure that we make it clear that anything that we cover in this module needs to be tailored to the legislation and policy and enforcement procedure in your jurisdiction. I believe that what we're going to be talking about today is generally applicable to different jurisdictions around the world, but if there is something that we're telling you that isn't consistent with what you're permitted to do, please make sure to follow your laws and standard operating procedures.

We would also encourage you to make sure you are familiar with your own agency's procedures and to draw on the learning available from the ICN, especially the Anti-Cartel Enforcement Manual compiled by the cartel working group, which is available on the ICN website in this area.

We hope that you'll find this presentation both interesting and helpful.

With that, I'm delighted to turn to Gary Bracken to begin our look at document review.

[ICN Training on Demand: Document Review]

[Gary Bracken, Assistant Director, Investigators and Intelligence UK Competition and Markets Authority]

GARY BRACKEN: Hello, I'm Gary Bracken, an Assistant Director in the Investigation Team at the CMA.

It might be stating the obvious, but it's important to have a good think about what it is you want to achieve out of a document review before you start it. One way to achieve this is to put a document review plan in place. Your document review plan should cover rules of evidence, keeping track of the review, objectives for the review, relevance, and organizing the evidence.

Rules of evidence: How will you meet the rules of evidence in your jurisdiction, for example, ensuring the rights of defense by confirming all material is recorded and reviewed, including for exculpatory documents; that you have procedures to deal with privileged material within your laws and rules -- and there's more on the issue of privilege in scene four, Collating the Evidence and Preparing for the Review -- that you have processes for how to handle irrelevant or duplicative documents; and that you handle material that is confidential or sensitive in accordance with your privacy and data protection laws.

Keeping track of the review: That is, how will you keep track of everything as you receive and begin to review it? For example, can you easily identify the material maybe through the use of unique reference numbers? Do you know where it came from? This is especially important in investigations where possession might be an indicator of culpability. For example, knowing

who owns a book containing notes of a cartel meeting and where it was found could be crucial to your investigation. Has it been reviewed, who has had access to it, and who needs to examine it?

Objectives for the review: Having established some basic foundations you can start to think about the objectives of the document review itself. At its core, the document review seeks to examine all the material gathered to determine how and in what way it's relevant to your investigation. For example, is it a key evidential document, whether inculpatory or exculpatory? Is it simply routine evidence or indeed irrelevant to the case?

Another way to look at the review objectives is to think about what it is you suspect is going on. What is your theory of harm, what do you have to establish, and then how are you going to do that?

Relevance: Before beginning the document review, those carrying it out should have a good understanding about what types of material may be relevant to the matters under investigation, bearing in mind material may be responsive to various aspects of a theory of harm. You should seek to give the review team guidance as to how they will find the evidence in practice and what to look out for. For example, which of the categories of material you've gathered is the evidence most likely to be found in, what evidence will have the greatest impact on your investigation, at what priority, and what are the resource implications?

These factors will invariably need to be tweaked as the document review progresses and you learn more about the case. Perhaps you can note

down any adjustments to the process as an annex to your document review plan. Also, think about how you'll ensure consistency in approach, for example, by encouraging reviewers to compare notes as they go along and to bring what the review finds to the attention of the entire team so that it can be factored into the overall investigation strategy maybe by way of daily catch-ups.

Organizing the evidence: Finally, looking forward, the document review plan needs to set out how you will record evidence when you find it and how you will organize it. For example, will you do it by theory of harm or by date or perhaps you'll need a mix of those. And remember to factor in cross-referencing for material that is relevant to more than one theory of harm, for example, evidence of price fixing and bid rigging.

Hello, Gary Bracken here again.

As we all know, the world is becoming increasingly digitalized and competition investigations are not an exception. To help us think about how to manage the transition, we're joined from Bogota by Juan Pablo Herrera Saavedra of the Superintendent of Industry and Commerce. Juan Pablo was responsible for the SIC's recent successful transition from a hard copy to digital case file.

Welcome, Juan Pablo. Thank you for taking the time to visit with us today. Given the very recent experience of your agency, we are particularly delighted that you are willing to share it with our audience today.

[Industria y Comerico Superintendencia]

[Transition from hard copy to a digital case file: an experience from

Colombian Competition Authority]

[Juan Pablo Herrera Saavedra, Deputy Superintendent for Competition Protection]

JUAN PABLO HERRERA SAAVEDRA: My name is Juan Pablo Herrera Saavedra, Deputy Superintendent for Competition Protection of the Superintendents of Industry and Commerce in Colombia.

With this contribution, the Colombian Competition Authority hopes to share how it successfully managed the transition from hard copy to a digital case file.

And let me start with a general overview. At the beginning of the year, the SIC was already planning to implement a pilot for the transition from hard copy to a digital case file. From March on, this plan was accelerated. The SIC had to cope with the challenges brought about, the confinement measures, particularly with the fact that our staff had to work remotely to ensure not only the continuity of our activities of inspections (inaudible) some control, but also to guarantee the rights to due process of all those being investigated.

The SIC took on the task of implementing a tool that would allow the virtual and remote access of those being investigated to the administrative files. To this end, the Office of the Deputy Superintendent for Competition Protection, working together with the Office of Technology and Information, implemented the Google Drive Enterprise tool to store, access, and share the digitized folders of the files corresponding to the investigations being carried out.

Additionally, with the purpose of orienting and guiding the

consultation through the affirmation tool, the SIC carried out two protocols aimed at both internal users of the entity and citizens. The latter was published in the webpage of the superintendents in order to be of common knowledge.

Today, it is gratifying to share with all the citizens that 100 of the active files on restrictive commercial practices against competition are completely digitized for the first time in the history of our agency.

In the digitalization process, a validation work of the information previously digitized was carried out before uploading to the drive. In total, 219 folders were digitized to complete the 712 folders of all administrative investigations carried out by the delegation for the protection of competition.

And the question is, how does the tool work? As mentioned before, the deputy superintendents created an application tool to effectively manage our current hard copy files and transition to our digital case files. With this tool, we got to create a system to efficiently index the information stored in each case file.

The tool works upon three main systems. The first system is a documental control files that were just a catalog. The second is a registry of physical and digital evidence, for example, like relevant documents for the investigation, as well as information that they've already collected in unannounced inspection from computer, mobile devices, and electronic mails. Finally, we developed a board where the bulletin of each case is displayed with the purpose of keeping the information updated.

As mentioned before, the estimated timing of the pilot project

became insufficient to keep up with the immediate needs posed by the COVID-19 pandemic. As you most certainly know, the pandemic situation directly impacted the regular exercise of activities and demanded from staff and contractors a quick process of adjustment and adaptation to the mobility restrictions and to the health and safety measures that were imposed by the government. For that reason, the Authority decided to prioritize the digitalization of case files and the acceleration of the implementation of digital files.

This situation led to the emergence of new practical challenges for the exercise of the Authority's functions. First, the Authority had to continue with investigations that were in progress, as well as to be able to take the new ones that resulted from possible violations to competition provisions in the context of the emergency. In this sense, there were only two possible solutions. First, to digitalize all the documents that were physically stored on site and, second, tocreate from scratch a mechanism of digital filing to keep track of the files from the new investigations that were taking place during the pandemic.

Nowadays, we are proud to say that the deputy superintendents has achieved the goal of digitalizing the 100 of our physical files, as I said. Also, all the investigations are working virtually and the files are totally digital. However, this process was not easy. And let me explain why. Let's talk about the process of digitalization of the files that implied for the staff of the SIC four key activities that you should consider for your own transition.

The first, careful digitalization of the physical information using scanners. The Deputy Superintendents began the process of digitalization, the

physical files of the administrative investigation, together with the Office of Technology and Information. This activity was carried out by 31 members of the team, including employees and contractors from the different working groups. During the digitalization process, a validation task was undertaken on the information that had previously been digitalized before uploading it to the drive.

Next, the authorities used the stored information in physical folders. These folders store several types of information containers, for instance, CDs, DVDs, USBs, hard drives, et cetera. So all of that information had to be downloaded and consolidated with the information that was scanned before. Next, finally, the staff of the SIC uploaded -- they consolidated information to the cloud, in this case, Google Drive Enterprise, and -- however, this is not as easy as it sounds. The staff of the SIC followed a strict protocol to handle information and further arrange it between public and classified.

Then, the SIC shared the digital case file with the parties under investigation, lawyers, complainants, and third parties. The digital case file was sent to the electronic mails of each one of the stakeholders. And, now, the question, how does the digital file actually work? Well, once the information is shared to the relevant stakeholders, they will have immediate access to the information depending on their status in the proceedings. Decisions regarding access to the case file by the parties had to be made previously by the case handler.

From the transition on, the information has been handled digitally. All the information has been received through our institutional electronic mail.

This new information is incorporated into the folders that are in the cloud. The staff in charge of the digital case file is responsible for keeping the information organized and updated, and so the catalog and the registry of physical and digital evidence.

Finally, let's talk about the benefits obtained from the this strategy of the digital file. The digital case file are now a reality. We can recognize today several benefits of the implementation of the digital files in our investigation. One of the most real ones is the possibility to have access to updated information immediately, any day at any time. Thanks to the implementation of the digital files, the procedure is swifter and the communication between the authority and the parties investigation are more efficient.

Finally, the parties in the investigation do not need to go to the Superintendents to obtain information related to the case. Also, they can take part of the investigation virtually.

Thank you very much.

[Industria y Comerico Superintendencia]

[Transition from hard copy to a digital case file: an experience from Colombian Competition Authority]

[Juan Pablo Herrera Saavedra, Deputy Superintendent for Competition Protection]

GARY BRACKEN: Thank you, Juan Pablo, for that helpful insight into some of the challenges you faced and solutions that you found, and many congratulations for delivering it all during the global pandemic.

[ICN Training on Demand: Document Review]

[Sophie Mitchell, Assistant Director, Antitrust UK Competition and Markets Authority]

SOPHIE MITCHELL: Hello, I'm Sophie Mitchell, an Assistant Director in the Antitrust Enforcement Team at the CMA.

By now, you'll be eager to start the actual document review and see whether you will find any evidence of an infringement. You've got your document review plan, you've got the information and material from the investigation on the file. Your file might even be fully digitized. What's left to do but start, right?

Well, no, it's not quite that straightforward. There are a few more things you should think about and do before you get started. Time invested now in collating the material and preparing for the review will save you a lot of work later on.

What exactly do we mean by collating the material and preparing for the review? Well, whether hard copy or digital, the information and material will ordinarily need some work before it is available in a form that you can review. Let me explain some options for doing that.

First off, it can be helpful to upload all the material received from different sources into one place. This can be as simple as a document folder or as complex as the use of an electronic document review platform, which many established competition authorities now use. Keeping material in one place will help you, first, with organizing the material and, second, with meeting your obligations as far as rights of defense are concerned, for example, by helping to

exclude privileged material or locate exculpatory evidence.

For hard copy material uploading to a document review platform might mean scanning or copying. Scanning obviously has a number of advantages. It effectively digitizes the material. It's a good idea to enable the optical character recognition or OCR function when scanning so that later searches can be conducted through the review platform. Scanning the material also means that more than one person on the team can examine it at the same time and this is really important. It keeps the original materials secure from loss or damage. But do consider how you will uniquely reference the material, for example, by using a code or number so that you know where it came from before you scan and upload it.

For digital material, uploading to a document review platform might require specialist processing and to that end you might want to reach out for some help from a data forensics team. Your agency might have such a team in-house or it might work with external data forensics specialists. They will be experts in backing up the material before you set off and processing and indexing it before it's ingested to the document review platform. They can also help you with preparatory work, like de-duplication and removal of embedded images that are incapable of being evidence, such as corporate headers and logos.

De-duplication avoids you reviewing identical material that is obtained from multiple sources, for example, an email copied to multiple internal addresses or an email discussing market sharing arrangements that is found during searches conducted at two competitors' premises. Removing images,

such as corporate headers or logos, which can have no evidential value can save valuable time during ingestion, but be careful not to remove attachments or other documents that, in the context of your investigation, may be potentially relevant evidence.

Most jurisdictions have strict rules and protections for material that may be privileged. It's worth making sure you have a good understanding of what the rules are where you are. Whatever the specific rules are, you'll probably need a procedure to make sure your review does not, by mistake, capture privileged material. Once again, you could note the procedure down as an annex to your document review plan so that you have a good record of what you did in case of future legal challenge.

The procedure can be as simple as agreeing on a list of privileged keywords with the businesses whose material you're searching, for example, the names of external lawyers or law firms, so that these can be removed before you begin your review. Again, your data forensics team can help you with removing any material that responds to the privileged keywords as part of your pre-review preparations.

And don't forget to include in your procedure what you're going to do with the material you remove. Not all material that responds to privileged keywords will, in fact, be privileged nor will it necessarily be relevant to the investigation. Will it be reviewed for privilege and relevance by an independent lawyer or simply returned to the business? However you decide to proceed, do keep a note of what you do.

Notwithstanding all your preparations, how you approach each review will be different, not least because the volume of materials you have will affect how to proceed. And, clearly, any approach will have to be flexible enough to fit the resources in terms of staff and the capabilities of your document review platform available to you. The sheer volume of material you're likely to receive will necessitate some form of filtering and sifting to reduce it to a manageable quantity. In practice, you may have to filter and sift a number of times to get to a realistic place to start the review.

Your data forensics team can likely help you with filtering and sifting using either a traditional document review platform or one with machine learning. Whatever your capacity, it is worth investing a bit of time to test the results you get to check everything is working as you expect and you're getting back what you want. Any filtering and shifting -- sifting should be determined by relevance as set out in your document review plan.

What will your relevance criteria be? Is there a date range that you can use? Is it emails between key individuals, is it key words and phrases or is it a mix of all three key? Key individuals' email addresses will certainly be found in the material you have, but consider whether you want all the emails where they're an addressee. Could you filter by recipient, sender, or copied? And do you want all their internal emails as well? Date ranges and the keywords and phrases, for example, names of relevant products, names of key individuals in the businesses under investigation or phrases, such as price increase, will probably be determined by your case knowledge.

As we mentioned earlier, you will find that you often have to filter and sift more than once, depending on how documents were gathered, to arrive at a pool of material that you can actually review. When determining how to change your approach, you could think about altering who's considered a key individual, reducing the date ranges, searching for a number of keywords that must appear together as part of a string rather than individually, or even where initial results are just too large, excluding a specific keyword entirely. Of course, when you get to examining material, you're going to have to make some decisions about its evidential value and, importantly, you'll need to record those decisions.

One way of approaching recording your decisions on an electronic document review platform is by tagging. Tags are a means of indicating which of your predetermined evidential categories -- yes, go back to your document review plan again -- the material corresponds to. Your data forensics team can help with setting up the tags for you. We will consider the use of tags on an electronic document review platform in the next part of this module.

When carrying out a document review, there are a few practical points to think about at the point of starting the review. First, think about how you are going to divide the document review exercise among different members of the case team. For instance, you might wish to have first line and second line reviewers with each potentially relevant document being reviewed by at least two members of the case team in order to build some consistency and quality assurance into the process.

How will you handle duplicate documents identified as part of the review? As mentioned earlier, an electronic review platform can filter out many of the duplicate documents before the case team begins the review, but even so, a certain number of duplicates or part duplicates will remain. This is particularly true for email correspondence where various copies of an email chain may exist in more or less complete forms. Removing partial duplicates can help to keep your case file to a manageable size, but it is important to ensure that relevant attachments to an email pathway down a chain are kept, along with the email they are attached to for context, if the attachments don't appear in later emails in the chain.

How will you record the results of your review? The use of tags on an electronic document review platform can be an effective and efficient way to keep a record of your decision-making on individual documents. Let's have a look at how tagging can work in practice. Once you have the documents loaded onto the electronic document review platform and you have set up your tags, sometimes referred to as codes, you are ready to start reviewing the documents. You'll definitely want to tag category for relevant material, probably with sub-tags to indicate hot material, those that are of particular evidential value or interest to your case, and the nature of the relevance or relevance details, for example, which elements of the theory of harm the material meets, remembering that there may be more than one, exculpatory material and not relevant material.

You might also want tags to record the businesses or contracts involved. Finally, you could need tags for privileged or duplicate material,

material requiring explanations, perhaps by way of witness interview, or those containing sensitive personal data, foreign language material, or where extra processing by your data forensics team is needed to enable you to view the material in the first place.

Whilst having a good spread of tags might appear helpful, too many can create future obstacles to effective searching and case administration. It can be helpful to focus on the key elements of the legal test, avoid too much detail and generally keep the tags relatively high level. It's also important to ensure that it is clear to all reviewers what the tags mean. A guidance document issued to reviewers with an explanation of the tags can help here.

Think also about which tags should be compulsory. For example, if a document is tagged not relevant, do reviewers need to apply any further tags? In the case of data sets containing a large number of irrelevant documents, the review can be speeded up significantly by reducing the number of tags that need to be applied to irrelevant documents. Conversely, where a document is tagged as relevant, is the reviewer also required to apply other tags such as those relating to the nature of the relevance or confirming whether or not the document contains sensitive personal data.

Many document review platforms allow you to set up conditional or compulsory tags so that a document does not show as fully coded until all the necessary tags have been applied.

[On-screen demonstration]

SOPHIE MITCHELL: An evidence matrix can be a useful way to

collate and organize those documents the case team has identified as being relevant in such a way that can be used in drafting internal papers or external decision documents, whether or not you are carrying out your document review on an electronic document with a new platform.

We are joined from Singapore by Serene Seet from the Competition and Consumer Commission of Singapore, who will talk about the CCCS's experience of using evidence matrices in document review.

[Serene Seet, Deputy Director, Legal & Enforcement Division, Competition and Consumer Division of Singapore]

SERENE SEET: At the Competition and Consumer Commission of Singapore, we work in case teams of between three to six people depending on the complexity of an investigation. Document and data review is a critical part of evidence gathering for the investigation. Once an investigation is completed, the evidence will form the substance of the CCCS's infringement decision.

We would like to share some of our experience in the way we conduct document review. CCCS investigation cases typically involve multiple parties and voluminous amounts of information gathered via leniency applications, dawn raids, witness statements, and information requests. To maximize the efficiency of our resources, strategizing any document review happens early.

First, we think about the theories of harm that could relate to conduct that we've been alerted to or that we have self-initiated as these theories of harm serve to target the categories of documents that we want to obtain. We

also consider the ways in which we can obtain such documents. It may be that we use a step approach so that we branch out gradually with our requests constantly re-evaluating the information we need to avoid document dumps.

Another way we structure our review is to use an evidence matrix. This can be in the form of a Word document or spreadsheet. The evidence matrix basically sets out the elements or nature of evidence required to prove the anti-competitive conduct based on the theories of harm and assists with organizing the evidence. This is important so every member of the case team will know what to focus on and the type of evidence to extract when carrying out the review of the information obtained. It also assists to summarize, sort, and organize the documents.

In cases of price fixing or information exchange, the matrix could be designed to populate the identification of the relevant conduct, such as pricing agreements and information exchanges. When distilling evidence of discussions between competitors, a table indicating the date the exchanges or communications took place, the forum in which they took place, the frequency, duration, and participants to the communication will be useful. It may also be useful to then organize the evidence chronologically or by types of meetings or by events arising from their conduct, for example, a price increase for a certain product or a price increase at a particular time and date.

The table could also include how the information was used, the impact of the agreement, and the parties affected. This could form much of the incriminatory evidence should investigations uncover it. Exculpatory evidence is

important, too, and another column could be incorporated into the evidence matrix to highlight such evidence. This will help the case team to assess the overall strength of the evidence and it is also useful to have another column or section of the table identifying possible gaps in the evidence. This will come in handy for the case team to obtain further evidence and close any gaps in the evidence to ensure that the evidence supporting the case is robust.

We hope this gives you an insight on what we think may be useful for case handlers when conducting an evidence review.

Thank you.

SOPHIE MITCHELL: Many thanks, Serene.

We have also found evidence matrices to be a useful way to organize the evidence.

For further reference, the ICN Training on Demand module on planning and conducting investigations provides further advice on how to organize the investigative file, prepare approved checklists, develop a case chronology, and so on.

[Chris Dodds, Assistant Director, Antitrust UK Competition and Markets Authority]

CHRIS DODDS: Hello, I'm Chris Dodds, an Assistant Director on Antitrust Investigations at the CMA.

In recent years, we've been hearing more and more about how machine learning can help(inaudible), such as speeding up analysis or identifying patterns within large data sets. One means by which machine learning can help

in document review is through predictive coding. We will also touch on continuous active learning a bit later. Both of these are relatively new tools that you may have access to through an electronic document review system.

To introduce us to this form of machine learning, how it works and how it could help you, we're joined from Washington, DC, by Tracy Greer from the Antitrust Division of the U.S. Department of Justice, DOJ. Tracy is an electronic discovery expert who trains and advises colleagues on how to get the best from this technology, for example, when agreeing on a predictive coding protocol that a party will use to sift material for relevance before production to the DOJ.

Welcome, Tracy. Thank you for taking the time to join us today. Given your extensive experience of this technology, we're really pleased that you're willing to share it with our audience today.

[Tracy Greer, Counsel, Antitrust Division, US Department of Justice]

TRACY GREER: Okay. So predictive coding, also known as technology assisted review, is a software tool available in document review platforms that can identify responsive documents. In more detail, it is an example of machine learning that, depending upon the version used, takes input from a subject matter expert on a subset of a document collection and applies that input to a larger collection.

There are a number of different products and services that we use throughout the day that use the same process, for example, email spam filters. When you first ask to receive a newsletter, those emails may be filtered out of

your inbox. But when I release them from my spam filter, the email service learns that those emails are not spam.

Predictive coding works the same way. The reviewer marks some documents as responsive and others as non-responsive, and then the algorithm can identify documents that are similar to each category and mark them as responsive or non-responsive. This is repeated until the document collection is classified. The point at which the process is finished is more nuanced, but for purposes of simplicity, let's stop here.

For what it's worth, I use the spam folder analogy because that was the first type of software used for these purposes. Many technology services have a similar approach. Product recommendations from an online retailer like Amazon, music choices on Spotify, but with predictive coding it is a binary choice, responsive/non-responsive, based exclusively on the input of the subject matter expert.

Predictive coding is built into some document review platforms. It's typically used to identify documents relevant to a document request in litigation. There are limitations to keep in mind. First, predictive coding works with text. It does not work for numbers, pictures, or similar things. The software typically identifies the documents without enough text for analysis and those documents are reviewed manually. Second, of particular relevance to this audience, the impact of the software when there are multiple foreign languages present is unclear. Most platforms claim that their flavor of predictive coding works with foreign languages, but it seems implausible to me that the software

can apply learning from German documents to Italian documents. I'm even more skeptical that this would work in a character-based languages like Japanese or Chinese.

The primary advantage of using predictive coding is to reduce the size of the document collection that has to be reviewed, which results in cost savings both to the producing and receiving party. Particularly in the United States, broad document demands routinely result in the production of hundreds of thousands of documents and often several -- several million. The Antitrust Division has neither the time nor the resources to conduct a page-by-page review of these productions. Therefore, it is in the interest of both sides to ensure that the production is no larger than necessary.

It is important the Division's experience suggests that productions made using predictive coding are richer than those made using manual review or using keyword searching. Richer isn't perfect, however. It contains more relevant and fewer irrelevant documents. It will not include only relevant and exclude no irrelevant document, which are called false positives, and will exclude some relevant documents, false negatives. Assessing the precise accuracy of predictive coding would require an extensive testing comparing productions using alternative production methods on the same document collection. Given the time and expense involved, it's not surprising that there has not been a great deal of real-world testing.

Even more difficult to assess is whether there are particular subjects for which predictive coding is not effective. For example, thinking of

cartel work, participants in a price fixing or bid rigging often understand that what they are doing is something they are not supposed to do. Participants may, therefore, use code words, communicate obliquely in ways that the algorithm does not identify. As a result, predictive coding should not be thought of as the only available tool in a document review platform. In the cartel example above, for example, document review platforms can identify communication patterns that would allow a cartel investigator to identify suspicious communications between competitors that should be investigated further.

The Division cannot require that parties use predictive coding. We encourage parties to do so and we continue to believe that it benefits the Division. We believe, however, that while predictive coding is useful to distinguishing between responsive and non-responsive documents, we do not believe that the algorithm is useful to identify especially relevant documents within a document collection. To do that, document review platforms use other analytical tools.

When receiving rolling productions, it is difficult to follow the formal iterative process of technology assisted review. The Division is having more success in utilizing specific figures, such as find more like these to locate similar documents, meaningful documents, and to cluster or categorize documents to help prioritize review.

I hope that this has been a useful discussion and an introduction to predictive coding.

CHRIS DODDS: Thank you, Tracy. Yes, that's been a great

introduction to predictive coding and evidence review, giving us some great tips about the capability and the limitations of this technology.

I'll just add a little to that, if I may, and then make a brief mention of continuous active learning.

With the CMA, we tried predictive coding in some recent antitrust investigations to help us prioritize our document review. We agree with Tracy that it can be a powerful tool under the right conditions, particularly where you have a very large data set of, say, several hundred thousand files. The process of teaching the system what is and isn't responsive so that it can reliably identify what you need does take some time. It's more useful the bigger your data set is. If it's less than, say, 100 000 documents, this may not be ideal.

Some electronic review platforms enable another form of TAR, continuous active learning. Again, the case team needs to manually code a sample of documents to teach the system what is and isn't responsive. Instead of tagging or coding up your material for you, it works by sorting your document set by predictive relevance, showing you the documents that are most likely to be relevant, but these teams can then review these and decide whether or not they are relevant.

The system then continues to learn and get better at prioritizing the remaining material as the review progresses. The team or party might stop the review when it appears that there is little or no relevant material left. You might want to consider continuous active learning if you want to get quickly started on the review using both people and machine learning together or perhaps

have a smaller data set.

We will now hear some examples from colleagues around the world of document review challenges you might face on your own investigations and how they overcame them.

First, we're joined from Pretoria by Mapato Ramagapa from the Competition Commission of South Africa, who will share the Competition Commission's experience of dealing with large data sets and handling highly confidential information on a significant market inquiry carried out by the Commission.

[Mapato Ramokgopa, Divisional Manager, Office of the Commissioner, Competition Commission of South Africa]

MAPATO RAMOKGOPA: Hello, everyone. My name is Mapato Ramakgopa from Competition Commission of South Africa, the CCSA.

It is my pleasure to briefly share with you our experience in dealing with large data sets and handling highly confidential information. The health market inquiry was initiated by the CCSA in 2014. This was in response to concerns regarding increasing costs of private health care in South Africa. The inquiry was a notoriously complex process involving extensive information and data collection across the healthcare markets. The data was collected from more than 248 stakeholders, which amounted to 545 gigabytes, representing the largest data set ever gathered by the CCSA and in the private healthcare market in South Africa.

The bulk of this data related to highly confidential and personal

information representing over 95 percent of medical transaction data. This includes patient records, diagnosis, treatment services done at the hospitals and other healthcare providers, billing and financial information, amongst others.

Given the confidential nature of healthcare transactional data, which by law cannot be disclosed, the inquiry had to develop a data de-identification tool and coding systems that would ensure the removal of personal identifiers in all the data sets. This process allowed stakeholders to provide the inquiry with data pertaining to patient information which ensured that patients' personal identities would not be identifiable while keeping each individual patient's records distinct for analytical purposes. The de-identification tool provided a hash key conversion on the patient's personal information into unique codes and developed key themes from the data, which made it easy to analyze records without compromising patient's details.

When the inquiry released an information report with findings and recommendations in 2018, several stakeholders requested access to data to interrogate the data relied on by the inquiry in its analysis. In formulating the appropriate framework, the inquiry engaged (inaudible) stakeholders, including the United Kingdom's Competition and Market Authority, from their experiences in the establishment of data management through a data room. The data was therefore accessed through the data room which provided external legal experts and other advisors to enable them the opportunity to run the (inaudible) verifications within a controlled environment. This was, however, subject to confidentiality undertakings.

A key part of success in the inquiry is the integrity in which the analysis was conducted and the way highly confidential and large data sets was handled in the technical analysis. This was a sensitive yet transparent and unbiased process.

Therefore, in conclusion, as competition authorities are now (inaudible) with enforcement issues and investigations in emerging markets, such as digital and online platforms, big data companies and industries, this experience is therefore available and the CCSA is more than happy to share our experiences.

Thank you.

CHRIS DODDS: Thank you, Mapato. I'm sure that colleagues around the world can continue to learn from the CCSA's experience in this area.

I'll now pass you over to Felipe Roquete from the Administrative Council for Economic Defence, CADE, in Brazil, who has some tips on how to improve and speed up the analysis of large volumes of digital evidence.

[Felipe Roquete, Conselho Administrativo de Defensa Economica, Brazil's Administrative Council for Economic Defense (CADE)]

FELIPE ROQUETE: We have had an interesting case here in Brazil, which we would like to share with you. During the investigation to a possible case of cartel in a procurement process, we were informed that criminal investigators from state prosecution services had carried out searches regarding the same matter. They had gathered about 80 terabytes of electronic data and apprehended laptops and several mobile devices. At first, we were interested to

have access to the digital evidence gathered by prosecution services. However, we soon realized that the criminal investigation team had a lot of knowledge about the case but did not have the operational means to process and analyze the material collected in its entirety. Thus, we devised a method to improve and speed up the analysis of the digital evidence.

Four of CADE's officials, case handlers, worked directly with the criminal investigators for a fortnight in order to devise the data analysis protocol that fits the specificities of the actual case, in other words, set apart any information about the most relevant individuals and companies from the rest of the data gathered during the criminal investigation, and allowed to focus on analyzing electronic data related only to the main targets of the investigation, that is, individuals and companies with more prominent involvement in the case. Then we mapped relevant keywords, dates, and corporate information, and applied it to the data analysis protocol in order to filter the information gathered from the electronic devices, excluding irrelevant sentences, for instance, thus reducing the volume of data to be analyzed from 80 terabytes to around 4 kilobytes.

CHRIS DODDS: Thank you, Felipe. And congratulations to CADE on managing the analysis of such a vast data set.

Finally, we wanted to share some of the CMA's experience of reviewing chat messages as part of a digital document review. Chat messages are increasingly a feature of the evidence base on antitrust investigations. Chat messaging services tend to be a faster and often a more informal communication

method which can be used across desktop, laptop, and mobile devices.

The records of these chats present their own particular challenges in evidence review, and so it's a good idea to plan how you will conduct your review of chat messages as part of your evidence review strategy. Consider how you will deal with the often informal or jargon-heavy nature of this kind of correspondence, are there particular acronyms you might look for or might you want to ask for a set of commonly used abbreviations alongside the messages themselves. This may be especially so for messages exchanged in chat platforms used by financial services businesses, which often employ shorthand terms.

Unlike emails, which tend to consist of shorter messages threads, individual chat messages can be grouped together in very long chat conversations when extracted to the digital e-platform, even where the messages are sent on different days and cover very different subjects. Chat conversations can take place between two or more individuals and can sometimes consist of hundreds or thousands of messages spanning a period of months or years. It will often contain a mixture of relevant and non-relevant messages and you'll need to decide whether to take the full conversation for your investigation file or any relevant groups of messages.

Messages sent on individual messaging platforms like Messenger, Signal, Whatsapp, Telegram are often extracted from mobile devices in two forms. First is a conversation which includes all messages exchanged between two or more individuals and, second, each individual chat message is extracted as an individual document. If extracted as a conversation, some review platforms

can present the information so it looks as it would have to the custodian. If so, this makes the material straightforward to digest and may speed up the review. However, be aware that the conversation view might not always show attached images and other message attachments, so you might need to review the individual messages in order to see these.

So when considering how many, if any, of the non-relevant chat messages from the chats to keep on the file, you might want to think about matters such as the volume of messages in a conversation, the proportion of relevant versus non-relevant messages, whether the non-relevant messages provide context for the relevant messages from an evidential point of view, and the nature of the non-relevant messages, such as whether they contain sensitive personal data.

Another issue you might have to consider is the time zone in which the messages were archived, which might be different to the time zone in which they were sent. For example, a message sent at 9:00 a.m. UTC, coordinated universal time, may have been archived in EST, Eastern standard time. It may show on the face of the document being sent at 5:00 a.m.

More generally, the types of information recorded by particular chat groups or chat rooms may differ. So you might want to consider what is the most evidential value when deciding on your approach to reviewing this.

ALEX SCHOFIELD: Thank you for joining us today. We hope you found this module useful and that it provided you with the foundations for a successful document review. It wouldn't have been possible without the help and

support from all of our contributors. So once again, many thanks to them.

We'd like to close with some advice and that is to expect the unexpected. Be flexible with your planning and be prepared to change your approach if something isn't working. And to point you one last time at the learning from the ICN's work, in particular, the other Training on Demand modules found in Series 6: Investigative Techniques and the Anti-Cartel Enforcement Manual, compiled by the cartel working group ,will help with document review, all of which are available on the ICN website. Do remember to check back from time to time for any new additions.

With that, we wish you good luck with your reviews and in finding the evidence you need for your investigations. Goodbye and good luck.