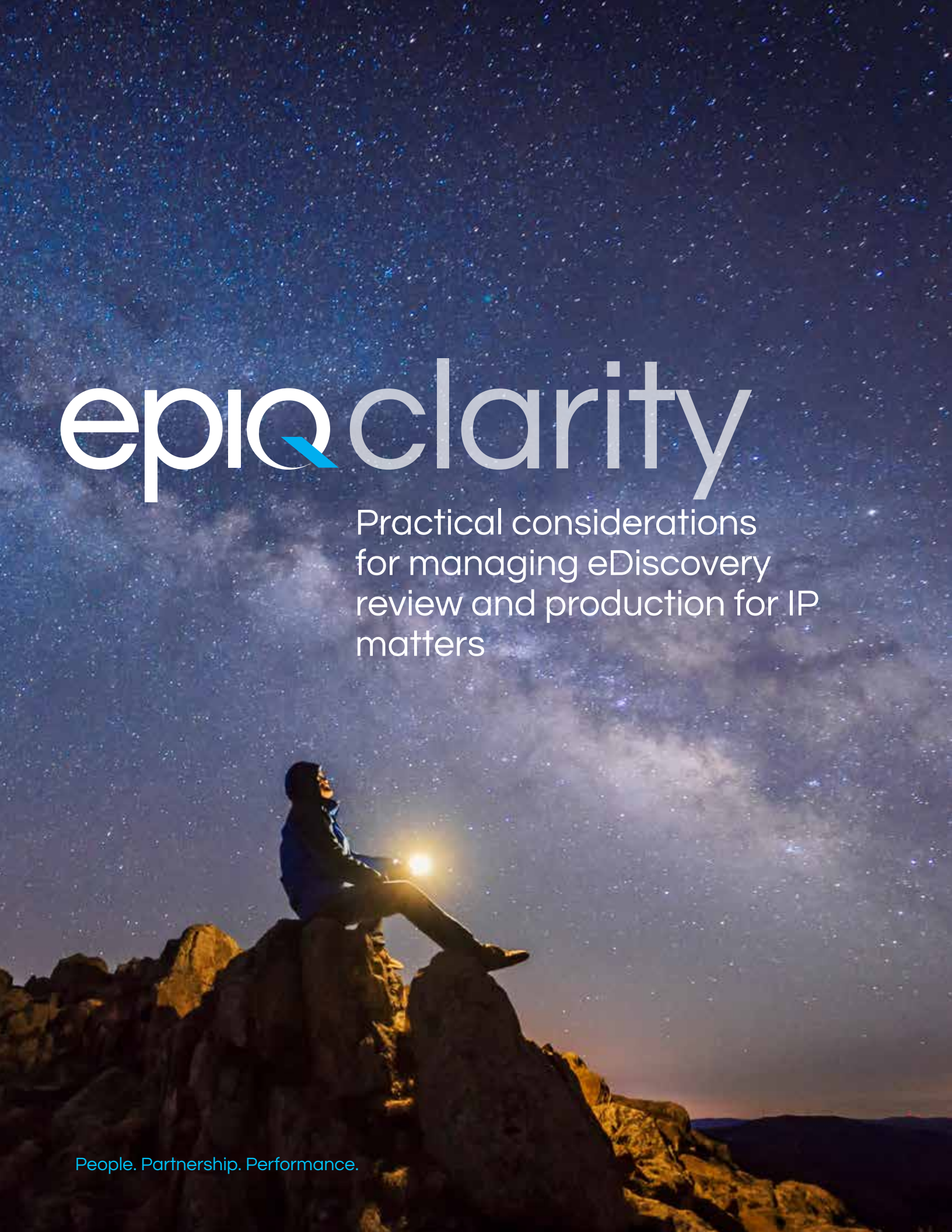
epiq clarity
Practical considerations for managing eDiscovery review and production for IP matters
People. Partnership. Performance.

Introduction

Most successful eDiscovery review and production workflows result from both a substantive and technical understanding of three factors:

- The contents of the electronically stored information (ESI) data set;
- the goals of the discovery process;
- and workflow options and limitations.

The purpose of this paper is to identify nuances and consistencies specific to intellectual property (IP) matters and to address methods of optimizing review and production solutions.

General data characteristics

a. Understanding your data

The first step to any successful eDiscovery review and production project is to develop an understanding of how your data set was defined. This includes understanding the scope of collection, the method of data identification for collection (e.g., by custodians, key terms, date range filter, etc.), and the content of the documents. Generalizations and hypotheses about document content can take a variety of factors into account, like participants involved in communications, types of interactions the participants would have, the general workings and processes of the company or entity at issue, and the context of the eDiscovery matter (e.g., the stage of litigation).

One of the best ways to develop a high-level understanding of the content of a data set is to identify categories of data within the set that you can extrapolate from general knowledge about IP litigation, as well as nuances of the data set at hand.

For example, a common data category in a pharmaceutical IP matter is a regulatory form required by the FDA. The FDA grants approvals for marketing or sale through an application process and there are different types of applications: a New Drug Application (NDA), an Abbreviated New Drug Application (ANDA), and an Investigational New Drug Application (IND). An ANDA application relies on previously filed or approved application information, and an IND requests permission to ship a drug prior to approval and is typically used for clinical trials.

The types of FDA applications included in the data set conform to a standard "virtual folder" organizational structure with specific data types and content based on that structure. During the processing, review and production stages of the electronic discovery reference model (EDRM), the embedded meaning of the foldering structure can be very helpful for understanding the purpose and context of the contents, as well as providing some insight into potential review methodology.

b. Understanding your limitations

Once you've achieved an understanding of your data contents, you will want to ascertain whether there are other factors that will limit your plans, or require adjustments. Likewise, you will want to identify and account for factors that will influence the goals and workflows for your IP matter. A basic consideration is the stage of litigation, especially in the context of the data set format and contents.

The inclusion of non-standard (e.g. non-MS Office, proprietary application, database or other transaction analysis type application, technical data formats, etc.) data types in your set is a common occurrence for IP matters. At the outset of your matter, communicating with your retained service provider or in-house eDiscovery team about whether your data set will, or is likely to, include non-standard data types provides an early opportunity for discussion regarding collection, processing, review and production in order to establish project design and workflow specifications. This will help mitigate processing issues and ensure all parties understand the scope and format of the available data and how it will progress through the workflow.

In the context of FDA submissions, examples of non-standard data types include: clinical study support files which may require conversion (e.g., SAS data formats), spreadsheets and data files which could require specific processing instructions to render

correctly (e.g., .xml, .dat, .csv, and .xsl file types), handwritten notes (often found in lab notebooks) and hard copies
for which the extracted text to be searched may not be accurately captured during the optical character recognition (OCR) process.

The size of the data files is important to consider as well, as large files (e.g. .xml and .ppt files, database application, and data format files) may be too large for tiffing (producing a file in a petrified format such as a tagged image file format a.k.a. TIFF), slow to render in an online review application viewer, or cause latency in the system during review. Where redactions are applied, large files may require customized tiffing protocols which result in a large volume of pages, sometimes in the tens of thousands of pages per document.

As an alternative to a standard tiffing approach to production, redacting and producing in native file format may be an option and can result in time and cost efficiency. Certain types of graphics files (e.g. CAD, SAS) may require custom viewers to be provisioned on the webserver or installed locally to allow a reviewer to view the content and to assign a review designation to the document. Additionally, if a data set contains mixed foreign language content, it may be helpful to utilize advanced analytics capabilities to automatically identify these files prior to review, so that they can be streamlined to a specialized review associate or translation application. Further, certain content like slide presentations, diagrams or flowcharts may require tiffing with color elements retained for legibility or redaction purposes, which may result in issues or latency with rendering due to the relatively larger file sizes.

c. Understanding your goals

Your data set and its limiting factors are the variables you'll consider when you define your goals for the matter. Identifying goals from the outset of the project will inform the workflows and tools you utilize, and will help to define whether the project is ultimately successful. It's relatively easy to identify a goal at a high level: who has not wanted their project to be more cost-efficient, or to complete a review in time for a deadline? However, you should only design a workflow around both a pertinent and realistic goal.

An ideal place to start when identifying pertinent and realistic goals for your project is often based on the stage of litigation. One example of defining a pertinent and realistic goal is if you are preparing data for production in response to a request for production (RFP). In this context, you will utilize the parameters imposed when identifying the data which is relevant for production, limitations on workflows utilized, and the format in which it must be produced.

For example, if you agreed with the opposing party about whether predictive coding technology could be implemented, that agreement will influence your workflow options. Milestone deadlines will also typically impact workflow and staffing consideration; if you must meet a short production deadline, you may need to utilize analytics options and ramp up with additional staffing accordingly. Alternatively, you may wish to perform a process commonly referred to as a 'data dump,' where you only have documents reviewed for privilege and personally identifiable information (PII).

Further, in situations where you are reviewing to understand the content of the data set, but need not designate coding categories for every document, you have an opportunity to leverage workflows that consolidate and minimize your review data set. The application and use of analytics tools and workflows which seek to identify and remove duplicative content can provide significant efficiencies and cost savings.

One example for a workflow of this nature leverages email thread identification to identify inclusive emails (the last-in-thread emails in a thread, or threads, plural, if they are split) and subsequently suppress non-inclusive emails (earlier-in-thread emails in a thread). The underlying rationale is that by reviewing only inclusive emails, reviewers will be able to view all content (as the content of non-inclusive emails will be contained within the inclusive emails) in a manner which is not duplicative. This type of workflow is a great example of factors involving the data set contents, its limitations, and the project goal coming together – just as the stage

of litigation influences how you may approach the review of a data set (in this example, the review would be consolidated and streamlined), it may also limit your capacity. For example, third party productions will often contain limited metadata, which may limit your ability to perform email thread identification.

Data analysis & organization

a. Categorizing your data

As discussed previously, categorizing your data set will help you to make sense of what you're working with and can be especially helpful if the data set is unfamiliar (e.g., from a third party). Think of the process that people who put together large puzzles often use: after first dumping out a box of 1,000 tiny pieces, a common technique is to start grouping them together by colors. It can often be easiest to group at a high level first: puzzles pieces which contain blue in one pile, pieces which contain green in another, and so on. From there, you can start sub-dividing categories: there may be pieces with green and brown which you can sub-divide in the 'green' category. During this process, it's most helpful to look at the picture on the puzzle box to guide your process – if the scene is of a forest, and almost all pieces have green in them, it may be helpful to divide the pieces into different shades of green. So, you want to take what you know (the picture on the puzzle box) and leverage it so that you can get the most out of your analysis (the organized piles of puzzle pieces).

Applying this analogy to the stage of litigation, if the litigation is advanced and there is no date cut-off that would preclude these types of documents, you are likely to come across deposition or hearing transcripts, exhibits, expert reports, and supporting materials. Another situation in which you are essentially taking unfamiliar (and often irrelevant) data and grouping it is when you are the recipient of a data dump. In this situation, utilizing analytics tools and workflows can help you group together similar content to provide a more organized, holistic perspective of the types of data included.

Depending on the review platform you use, there are a variety of options when it comes to analytics tools.

As previously discussed, predictive coding (sometimes called technology-assisted review or "TAR") is an increasingly common approach to leveraging technology to improve productivity and efficiency. Analytics tools that leverage structured analytics (like email threading and duplicate identification) can help sort your data set in a manner which provides for workflows which can consolidate content, efficient and effective targeted searching and data organization, and present documents in a way that allows reviewers to understand the context of communications and increase the consistency and efficiency of their coding. Conceptual analytics tools can group and sort your data, identify similar data, and identify patterns and inconsistencies by identifying consistent traits ("concepts") in your documents, and how these concepts inter-relate within your data set.

b. Identifying your data set

The converse of pouring through an unfamiliar data set, is when you are the one who is identifying the review set to begin with (e.g., when you will be the producing party). Then, you are in the best position to understand its contents and to develop parameters which ensure you are reviewing the documents which are most likely to be relevant.

One of the most common ways to define a data set is by applying keyword term filters, and there are many factors to consider when creating or revising a set of search terms. Different custodians (in different roles) may use varied language to discuss the same concept – different jargon should be taken into account when developing terms. For example, a member of the product development department may use different language than an executive officer would. Product names and components are often abbreviated, and code words are often used and may vary across company divisions or even individuals. Additionally, products may be described differently over time, through the use of different code names or with the development of different versions or iterations of a product. It is important to recognize that complex technical terms may be more likely to contain typos or misspellings. In addition to filtering using keyword terms, date filters are often applied for patent matters, as there will be specific dates that are relevant to the development of the product.

A common tactical approach in IP litigation is to provide the opposing party with a production data dump. In preparing this type of production, analytics tools will provide you with additional means to filter privileged, personally identifiable information (PII), private health information (PHI) or sensitive business information (SBI) content. Conversely, in those situations where you're on the receiving end and are parsing through either a data dump or large third party production, analytics may facilitate the ability to mitigate the size of the data while retaining the substance by leveraging email threading values and duplicate identification or concept clustering. The ability to use analytics may depend on how the parties have agreed to produce their data (and any controlling agreements regarding the use of analytics), the types of documents, or
the availability or quality of extracted or OCR text. This relates back to the meet & confer and any ESI agreements between the parties.

c. Reducing and organizing your data set

Once you understand the categories of data your set contains, or is likely to contain, you can search for these categories and leverage them in your workflows. In addition to special considerations for collections and processing, identifying these sets may provide you with the opportunity to cull or bulk assign coding designations (aka, 'bulk coding') for these categories. In addition to categories of data previously discussed, more examples of categories common in pharmaceutical IP litigation include: Medwatch forms, as part of the adverse event reporting (AER) process; marketing support documentation, like product labeling and packaging; certified patents and file histories; and published studies.

Specific types of data pertinent to software or medical products manufacture IP matters (e.g. system files, source code, schematics, etc.), may need to be segregated (removed from the general review population) for specialized review by subject matter experts, isolated or restricted from general access due to protective order, privilege or IP, culled (excluded for irrelevance), or bulk coded. Understanding the company structure and style of specific, consistently formatted communications (like forums or mailing lists) is helpful, as this information could be culled or bulk coded and can also be leveraged across matters. Like any matter, analytics tools can be a helpful resource to identify these categories if they are unknown or not easily searchable.

d. Preparing the protocol and coding panel

Understanding the content of your data set will provide you with a blueprint for what you want reviewers to code during first level review. For example, you may want to internally discuss whether the set will contain SBI, whether reviewers should code documents containing SBI, and if SBI should be redacted if included in a document with responsive content. It will also be helpful to distinguish between SBI that is responsive (e.g., source code related to the product at issue), and that is not responsive (e.g. un-related or not at issue product information).

Data may also require redactions for privilege, private health information (covered by HIPAA), and other personally identifiable information (e.g., social security number, personal bank account information, personal address, personal phone number, personal email address, etc.). Additionally, there may also be multi-level confidentiality coding or specific wording which should be used pursuant to a protective order.

Finally, it may be beneficial to prepare coding that allows reviewers to designate situations in which they cannot code documents, like further review coding or choices for technical issues and foreign language.

e. Performing organized and targeted batching

Organizing your data before batching and prioritizing certain categories will help you realize the largest return on your investment during the course of review. Batching by specific custodians, or groups of custodians in the same role, helps keep jargon consistent. Chronological batching helps reviewers understand product development and how events unfolded. Specific batching techniques, like batching high-level PowerPoint file types and similar documentation which could provide substantive knowledge, can help educate reviewers early in the review process while providing content that is likely to be responsive.

Batching specific categories of documents (like source code) to reviewers with specialized knowledge or training helps ensure consistent, accurate coding. Patent matters tend to have large, technical documents which often cause technical issues. Identifying, batching and assigning these documents to a specific reviewer or team lead mitigates the impact of these items on review efficiency as well as ensuring consistent coding.

f. Considering the makeup of your team

There are also considerations to take into account when staffing your review. Utilizing reviewers with IP review experience is helpful as there typically are consistent substantive traits across these types of matters, e.g., defenses (attacking validity of patent, independent development, reverse engineering), damages (often times financial documentation or reports). Incorporating reviewers onto your team who are familiar with the litigation of specific company is helpful not only for a specific matter, but across matters, because the reviewer(s) carry institutional knowledge they can leverage and also help educate less familiar colleagues.

g. Preparing reviewer resources

As with every eDiscovery task, when you cultivate and share knowledge from the beginning of a case lifecycle you increase the efficiency and effectiveness of team collaboration. Insight gained during search term development can also be integrated into resources for reviewers such as a glossary (especially helpful with the technical terms, code names, and abbreviations which often appear in IP matters), a code name list, a timeline, etc. Understanding terms related to other products or versions not at issue helps reviewers identify documents which may need SBI redactions or are irrelevant (persistent highlighting helps here as well). Privilege/PHI/PII filters and highlighting may also be helpful. Exemplar documents are especially helpful when introduced during review training for IP litigation (a training given by counsel who is typically entrenched in this information, and is given to a group of reviewers who are learning voluminous technical details and terminology in a short amount of time), as these help put the protocol into content and help the reviewers learn how this information is actually depicted in documents. Pictures and diagrams also help calibrate reviewer's understanding.

Beyond first level review

a. Performing quality control

One of the most integral parts of the eDiscovery review process is a quality control (QC) workflow. This is the point where the diligence that you performed identifying and organizing the data set will especially pay off: the categories and patterns you discovered can easily be leveraged to ensure the review team is coding documents consistently.

b. Institutional knowledge

Identifying, creating, and leveraging institutional knowledge across matters allows you to partner with your vendor and review team to add value. Documentation is one of the key aspects of retaining information for later use, and creating documents like an issue log, search term tracker, glossary, key party list, or timeline will allow you to leverage that same information on later iterations of your current case or on related matters. Identifying workflows, tools and processes that allowed you to increase your review efficiency and effectiveness will allow you to utilize the same or similar techniques in like situations. Conversely, identifying pain points and issues can help you proactively ensure the same mistakes are not made in the future.

V. Conclusion

We have seen that unique and non-standard data types require particular consideration to mitigate processing, review or production issues. We have also seen that the considered development and use of search terms for culling along with the effective use of TAR and other review optimization tools and workflows can assist with reducing review volumes while increasing review efficiency and consistency. In conclusion, effective planning, communication, documentation, flexibility to adjust to workflow limitations and end goal risk mitigation are key tenets to effective IP eDiscovery project management.