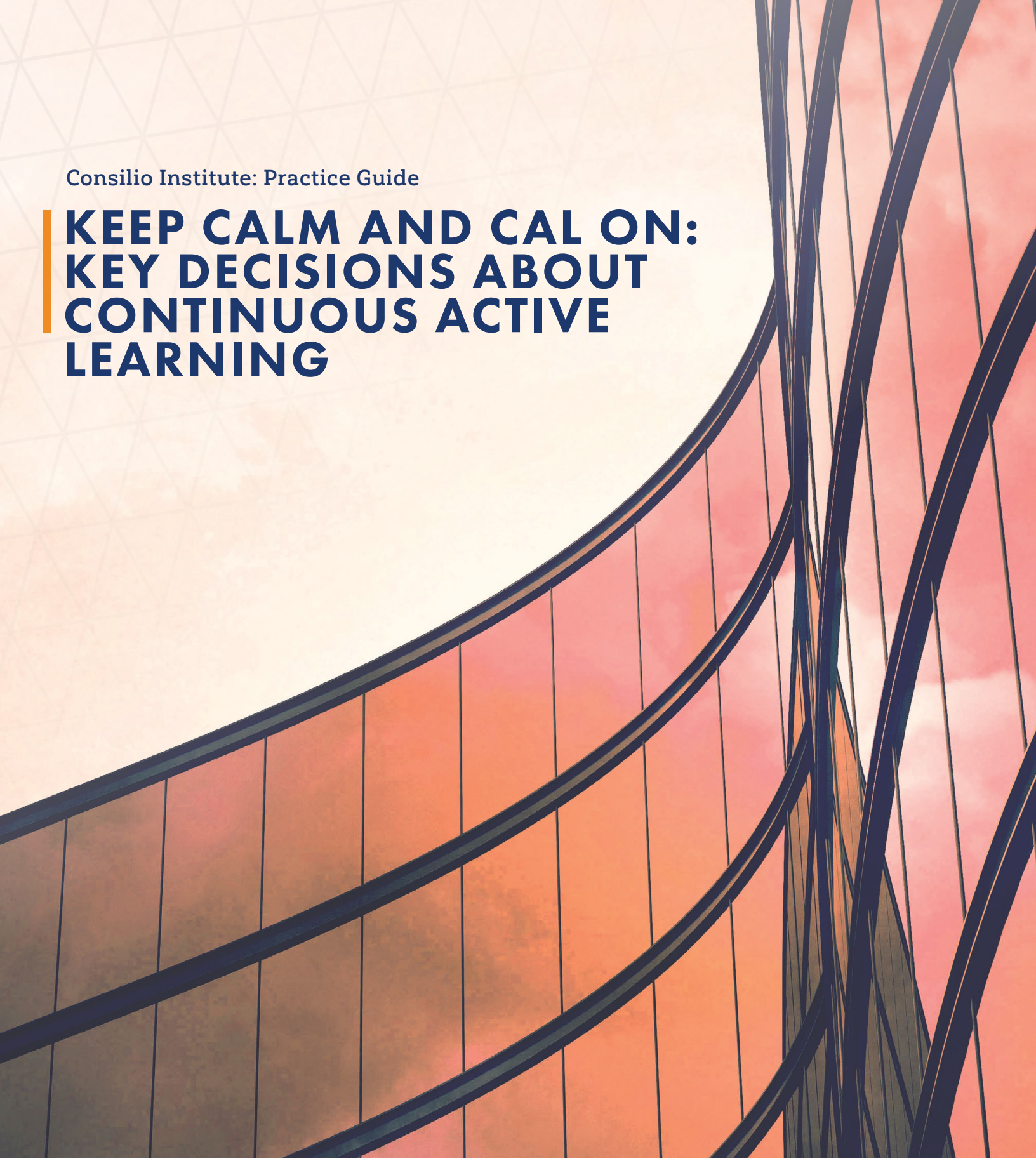

Consilio Institute: Practice Guide

KEEP CALM AND CAL ON: KEY DECISIONS ABOUT CONTINUOUS ACTIVE LEARNING

Xavier Diokno
Senior Director – Data Analytics / Innovation, Consilio

Consilio / ADVANCED LEARNING INSTITUTE

Consilio / ADVANCED LEARNING INSTITUTE

Consilio Institute: Practice Guide

KEEP CALM AND CAL ON: KEY DECISIONS ABOUT CONTINUOUS ACTIVE LEARNING*

CONTENTS

*Adapted from the webinar "Keep Calm and CAL On," co-presented by Xavier Diokno and Cleary Gottlieb Managing Discovery Attorney Michael W. Bohner on March 23, 2022, available at https://www.consilio.com/resource/webinar-keep-calm-and-cal-on/.

Consilio / ADVANCED LEARNING INSTITUTE

KEEP CALM AND CAL ON: KEY DECISIONS ABOUT CONTINUOUS ACTIVE LEARNING

Introduction

On any new matter, an eDiscovery professional must assess the project's goals and whether analytic tools can help achieve those goals. With the increased use of Continuous Active Learning (CAL), eDiscovery professionals must also determine whether their project is a good fit for CAL.

CAL is a workflow that uses a machine learning application to help identify responsive documents. Documents that are reviewed and coded for relevance are submitted to the CAL application for analysis. The application develops a model (or "classifier") by training on the responsive and non-responsive content. As training continues, the CAL model begins to rank documents based on their level of responsiveness. These ranks can then be used for a variety of purposes, such as prioritized review, quality control, and culling or excluding documents from review.

At minimum, the five steps required to setup and leverage a CAL workflow are:

1. Identifying the documents that CAL will analyze
2. Creating the CAL project (some applications refer to this as an index or build)
3. "Jump-starting" the CAL training with an initial set of training documents
4. Reviewing documents, which CAL ranks as responsive
5. Continuously updating CAL with reviewed documents

This practice guide will review the key factors that must be considered when facing the three main decision points related to leveraging CAL: deciding whether to use CAL, deciding how to use CAL, and deciding when your CAL process is complete.

Deciding Whether to Use CAL

Deciding whether to use CAL on a particular matter requires weighing several factors. These include the project's goals, the document volume and composition, the available time, the potential costs, and other considerations.

Project Goals

Document review efficiency and saving costs is a priority on almost all projects. Typically, the most expensive or "wasted" expense is when an attorney reviews a non-relevant document. CAL can reduce this expense by "bubbling-up" or prioritizing relevant documents. Moreover, at some point in the review, CAL will have prioritized most, if not all the responsive documents. Any remaining unreviewed documents are likely non-responsive and can be potentially excluded from review.

Performing an investigation and quickly finding the most important or key documents can also be a project goal. These projects are often related to a specific event or issue and involve gaining an understanding of the data before an official request for production. In these cases, specific search terms, date ranges, and custodians can be used to limit the document set. Moreover, a two or three attorney team can limit their review to the documents CAL prioritizes. As important documents are found, CAL will use these documents to find other documents containing similar content. This allows a smaller team to review a smaller subset (e.g., 500 to 1,000 documents), instead of a large document set (e.g., 5,000 to 7,000 documents).

In addition to quickly identifying key documents, CAL can enhance quality control. Documents available for QC are those where the reviewer's coding disagrees with the CAL rank. These disagreements can be

identified by searching for documents that have high CAL ranks (highly responsive) and were coded as non-responsive as well as those that have low CAL ranks (highly non-responsive) and were coded responsive.

Volume and Composition

CAL can also be applied to varying project sizes. This can range from a large document review (e.g., 400,000 documents and 20+ reviewers) to a small discovery exercise (e.g., 3,000 documents and two reviewers). One of the few limitations of CAL is that the documents must be suitable for CAL analysis. As a best practice, analyzing the documents prior to running CAL and excluding documents that have little textual value will make the CAL process more efficient and will reduce the number of documents sent to the CAL application. These documents may include images, audio files, system files, or documents that were created by a proprietary application (e.g., engineering software).

Timing

Timing is also a factor in deciding whether to use CAL. Ideally, CAL should be used at the beginning of the project, allowing you to take advantage of CAL ranking early on. However, CAL can also be applied after a project has started. In this case, any documents that were reviewed can be used to train CAL, allowing you to prioritize the remaining documents and to QC reviewed documents.

Other timing considerations include whether any reporting of the CAL process is required by the requesting party, the level of reporting needed, and the timing of those reports. Moreover, there is additional overhead if CAL is used to cull non-relevant documents from review. This may result in time spent prior to the start of review negotiating a CAL protocol with the requesting party. Validating the CAL results, which often includes sampling the documents to be excluded, also adds time to the process.

Costs

The cost of running CAL should also be considered. CAL may be charged in various ways, depending on the size of the project, who is performing the document review, and the level of support needed. CAL is often charged according to the number of documents sent to the CAL application. Moreover, in cases where a review team is needed, a per document or hourly cost may be incurred.

When CAL is used to cull documents from review, the cost savings are generally greater on larger projects, since there is a larger volume of documents to be culled. Before deciding whether to run CAL, it's good practice to analyze the document set, including estimating richness. Based on this analysis, a cost estimate can be calculated depending on the workflow selected: linear review of the full document set, a CAL prioritized review across the full document set, or a CAL prioritized review limited to responsive documents.

Negotiation

Finally, the ability to negotiate a CAL protocol and the time needed to negotiate with the requesting party should be considered. Negotiating the use of CAL is generally advised in cases where CAL will be used to cull documents from review. The negotiation process involves describing the CAL workflow in the ESI protocol and negotiating the protocol with the requesting party.

Deciding How to Use CAL

CAL can be used in a variety of ways ranging from large scale reviews to smaller investigations. Moreover, today's CAL applications offer several features, making it easier to train CAL models.

The most common use case are relevance reviews that involve (1) having CAL train on coded documents and (2) prioritizing high ranking documents for review. Another good use of CAL is to find important documents, where CAL is trained on key documents rather than broader responsive documents. After CAL

trains on enough key examples, a review of high-ranking documents should uncover more important documents.

Documentation

Once CAL is selected for a project, it is important to document the CAL process. Documenting the CAL process is critical when negotiating its use with an opposing party. The ESI protocol should include a description of how CAL will be used, an overview of the workflow, and the metrics that will be monitored and reported. In cases where CAL will be used to cull documents from review, it is advisable to notify the requesting party early in the discovery process and allows for the CAL review to start while negotiations are ongoing. Moreover, drafting an effective review protocol that clearly delineates between responsive and non-responsive content should reduce the number of miscoded training examples and increase CAL's efficiency.

Initiating the CAL Workflow

The CAL workflow starts by identifying the documents to be analyzed. Culling the document set using the appropriate methods is next. This may include the application of search terms and date filters, exclusion of specific file types, and email threading. Remaining documents are screened for system files, junk files, etc., as well as documents that have too little or too much text (e.g., text size is greater than 5MB). Documents that are screened-out should be analyzed and reviewed if necessary.

The remaining documents are submitted to the CAL application. The application processes each document by extracting its features and characteristics, and then indexing the document. This process creates a logical filing system that allows the CAL application to query documents based on their features. (e.g., find all documents discussing competitor analysis).

The next step is to train CAL using an initial set of responsive and non-responsive examples. These documents are usually either a random sample or documents that have already been coded. The random sample is typically based on a statistical probability, such as a 95% confidence level with a +/- 5% margin of error, which equates to approximately 384 documents. A person that is familiar with the matter and issues should review and code the sample.

The completed sample is then used as the first set of training samples to "jump-start" CAL's learning. Using the documents' features and coding values, CAL begins ranking documents with similar features. In addition to training CAL, random sample results can be used to estimate the number of responsive documents in the population. For example, if 95 documents were found responsive within a 384-document random sample, an estimated 25% of the documents are responsive.

Likewise, a seed set can also start CAL training. Seed set documents can be any previously coded documents that can assist the CAL training. These are documents that may have come from earlier custodian interviews or investigations. A seed set can also be created by reviewing a small set of documents that are based on targeted keywords and custodians. In cases where CAL is used for culling, using a seed set may increase scrutiny from a requesting party, including possible disclosure of these documents.

CAL applications may also offer features to enhance CAL training. Synthetic or fictitious documents are created by the project team as hypothetical examples of the kinds of topics or materials being sought. For example, if emails related to a product defect were considered important, a synthetic document could include descriptions of the client's product, a description of the defect, including when and how it was identified, and any reaction by the client after the defect was known.

Pre-Trained or Transferable Models

Some CAL applications also feature pre-trained or transferable models. Pre-trained models are developed by the CAL application vendor and can be used "out of the box" on a new CAL project. These models have been trained using documents and concepts related to a specific subject matter. When applied to a new set of documents, the model ranks documents according to the similarity of the subject matter used in training the model. For example, if a project is focused on identifying documents showing prejudicial treatment, a discrimination model may assist in identifying these documents. Moreover, CAL applications may include several pre-trained models that cover various subject matters. Depending on the project's focus, one or more models can be used to start CAL training.

Transferable or portable models allow you to save the model at the end of a CAL project. This feature allows users to save their work product in the form of a trained model. When a new matter arises, the saved models can be applied to the new matter. For example, if a CAL model was used in identifying documents related to sexual harassment, the CAL model can be saved and then later applied on future projects related to sexual harassment.

Process Automation

Many CAL applications also automate prioritized review by automatically creating review batches or assignments. After CAL ranks documents, the CAL application automatically identifies the top-ranking documents and uses these to create review batches. The CAL application then monitors the review team's progress as they start reviewing documents. When the team starts running low on documents, the CAL model gets updated with newly coded documents and the documents are re-ranked. This process repeats with the automatic creation of new batches containing the most recent top-ranking unreviewed documents.

To reduce the amount of training needed, many CAL applications can also intelligently identify training documents for review. CAL will analyze several factors in selecting training documents, including documents that contain concepts that have not been trained on, documents that form a large group of contextually similar documents, as well documents that are considered borderline responsive or uncertain. In addition to a random sample and seed set, training on documents selected by CAL early in the CAL process can speed up CAL's learning.

Parallel Processes

While CAL runs in the background, any processes or methods used in a traditional first pass review can still be applied in a CAL review. These may include sampling and QC, as well as processes for handling special documents, such as key documents, privileged documents, and documents containing private or confidential information.

Sampling and QCing reviewer decisions early on and throughout the review workflow is also important. At the start of review, the reviewers are likely still familiarizing themselves with the document subject matter. Moreover, they may not have a clear understanding of responsiveness. These issues are compounded when using a large review team. To identify mis-coded documents, samples should be taken across different ranges of the CAL rankings (low, mid, and high-ranking documents). Reviewers can then be provided with guidance and feedback based on the QC findings.

Since most responsive documents will be reviewed in a CAL workflow, reviewers can still tag documents

containing privileged, private, or confidential content. Workflows used to handle these types of documents in a linear review can still be applied in a CAL review. Moreover, as described earlier, some CAL applications can run multiple CAL models. One option would be to have a responsive model that is trained using responsive and non-responsive documents and a privilege model that is trained using privileged and non-privileged documents.

Deciding When You Have Completed CAL

On large-scale review projects, CAL can significantly reduce review costs by limiting the number of documents reviewed. Since the premise of CAL is to prioritize and review high-ranking responsive documents, there will come a point when review can stop because the remaining unreviewed documents are overwhelmingly non-responsive. This section describes methods for determining when review can stop, including evaluating key metrics and performing an elusion test.

Each method can be used as a building block to show that CAL was used in a reasonable and defensible manner. As described earlier, the first steps to support a defensible CAL workflow are (1) providing early notice to the requesting party and (2) documenting the CAL process, including any relevant metrics and validation methods.

Evaluating Key Metrics

Evaluating richness can help determine when CAL is complete. As described earlier, a richness estimate represents the total number of responsive documents in the CAL population (documents submitted to the CAL application). The random sample used to start CAL training also serves as an initial richness estimate. Multiple richness samples should be completed throughout the CAL review to assess whether CAL is complete at a given stage. Determining whether CAL is complete can be based on the percentage of the total number of responsive documents found throughout the CAL review against the richness estimate (estimated number of responsive documents in the population).

In addition to richness, the responsive rate (or found rate) can indicate whether CAL is complete. This rate represents the percentage of responsive documents that reviewers are finding across the most recent review batches. At the start of a CAL review, the responsive rate is typically high because CAL pushes responsive documents to the front of review. A high responsive rate at the start of CAL is usually an indication that CAL has accurately ranked highly responsive documents.

As review progresses, the responsive rate begins to decrease (in some cases drastically), possibly plateauing or bottoming out towards the later stages of review. A responsive rate that continues to stay low is generally an indication that review is complete. When the rate stays low, an analysis of the most recent group of responsive documents found (e.g., 50-75 documents) should also show that these documents are marginally responsive and contain no critical or new information.

Another helpful metric to monitor is the current CAL rank. CAL applications rank documents within a fixed range. For example, highly responsive documents are assigned a high rank (or score), such as 1.00, whereas highly non-responsive documents are assigned a low rank, such as 0.00. The CAL rank can be monitored to help determine whether review is complete. In most cases, a very low rank is an indication that most, if not

all, of the responsive documents have been reviewed. Any remaining unreviewed documents are likely non-responsive, an indication that review can stop.

Elusion Testing

In addition to monitoring the above metrics, an elusion test can be performed to validate whether CAL is complete. An elusion test estimates the number of responsive documents that would be left behind if review were to stop. The point where review stops is defined by the lowest CAL rank reviewed (e.g., all documents ranked above 0.20 have been reviewed). This "cut-off" rank (or score) defines the population of documents that will be culled from review (e.g., all documents ranked 0.20 and below will not be reviewed). The test can be performed multiple times throughout review to assess whether CAL is complete.

The test starts by reviewing a random sample that is drawn from the documents that have not been reviewed. The elusion rate is then obtained by calculating the percentage of responsive documents found in the sample. The estimated number of eluded documents (number of responsive documents left behind) is calculated by applying the elusion rate to the number of unreviewed documents. Many of today's CAL applications now automate these steps.

The elusion rate can also be used to estimate recall. Recall is the percentage of responsive documents found (throughout the CAL review) out of the total number of estimated responsive documents in the CAL population (documents submitted to CAL). A recall estimate is often requested by the opposing party. Despite the lack of an industry standard, a reasonable recall rate typically falls in the 70 to 80% range.

Conclusion

On any new matter, an eDiscovery professional must assess the project's goals and whether CAL can be leveraged to help achieve those goals. When considering CAL, there are three main decision points to face: deciding whether to use CAL, deciding how to use CAL, and deciding when your CAL process is complete.

- Deciding whether to use CAL requires consideration of several high-level factors such as the project's goals, data volumes and types, costs, timing, and negotiating posture.
- Deciding how to use CAL requires consideration of process documentation, initial training methods, whether and how to leverage automation and transferable models, and incorporation with document review processes.
- Finally, deciding when your CAL process is complete requires consideration of how best to evaluate key metrics and whether to perform elusion testing.

ABOUT THE AUTHOR

Prior to becoming an attorney, Xavier worked in the information technology industry for several years in database administration, telecom and VoIP implementation, and software development. As a senior director of Consilio's Data Analytics group, Xavier oversees projects that involve, Technology-Assisted Review, Immediate Case Assessments™, and analytics research.

Xavier Diokno has a bachelor's degree in computer science from Southern Illinois University, a master's degree in computer science from the University of Illinois at Chicago and a juris doctor degree from DePaul University College of Law. He is licensed to practice in the state of Illinois and the United States Patent and Trademark Office. Xavier oversees Consilio's Data Analytics team, where he advises clients on how to leverage technology in supporting their projects.

Xavier Diokno
Senior Director, Data Analytics

m +1.312.638.3130
e xdiokno@consilio.com

consilio.com