FIT FOR PURPOSE:

THE BEST OF EDISCOVERY TECHNOLOGY TODAY

CDS

THE HYPOTHETICAL 4
RELATIVITY'S APPROACH 5
COLLECTING DATA 5
UPLOADING DATA 6
REVIEWING DATA 8
TRANSLATING DOCUMENTS 10
REDACTING PII 11

DISCO'S APPROACH 12
COLLECTING DATA 12
UPLOADING DATA 13
REVIEWING DATA 15
TRANSLATING DOCUMENTS 17
REDACTING PII 19

APPENDIX 20
BIOS 20
THE HYPOTHETICAL, EXPANDED 21

eDiscovery practitioners today have many sophisticated tools to choose from that capably handle data at every stage of the EDRM. Legal teams can opt to select technology that best fits the case and the data, as well their specific processes and workflows.

While more flexibility is beneficial, evaluating platforms can be challenging. CDS provides full support for leading eDiscovery platforms Relativity - both Relativity server and RelativityOne - and DISCO, as well as 20+ other eDiscovery technologies. Our team of legal-technologists helps clients make informed decisions and select the solution/s that fit their needs. It's rare to have an apples-to-apples comparison to identify advantages or drawbacks of one tool over another.

For PIIP 2023, CDS' annual legal data management conference, we developed a hypothetical real-world scenario and asked eDiscovery software providers Relativity and DISCO to illustrate how their technology addresses the predefined challenges, empowering best practices, streamlining workflows, and improving efficiency.

THE HYPOTHETICAL

FixIn You Pharmaceuticals was on the brink of launching their first dynamic blood transfusion product, BloodFast, when DoubleTheFun Pharma launched a nearly identical competitive product. FixIn You believes that former employees working on their BloodFast product team stole research and trade secrets, using them to develop DoubleTheFun's competing product. Plaintiff FixIn You is seeking injunctive relief and damages against Defendant DoubleTheFun.

The plaintiffs have collected the BloodFast product team's communications, including the former employees, which consists of 17 custodians residing in the US, Italy, and the Netherlands. Data sources include Microsoft 365 Exchange, Teams, Excel, Word and SharePoint data, in addition to company issued devices which include iPhone 13s, iPads, and Slack totaling 1.65 terabytes of data.

Counsel for the plaintiffs need software solutions to do the following:

- Determine the source of leaked or stolen BloodFast information in the 1.65 terabyte dataset.
- Analyze the defendant's productions to confirm the source of the leaked information.
- Help identify communications by the defendants in which they knowingly accepted and used stolen proprietary information from the plaintiff's former employees.

Our software partners presented solutions that could support content spanning over 20 years and nearly 370,000 documents for processing. Additionally, special considerations need to be made for foreign language ID and translations, PII and PHI identification and redaction, and short message data handling.

The complete hypothetical and assignment can be found on page 21.

RELATIVITY'S APPROACH

According to **Shamir Colloff,** senior product manager at Relativity, the first step would be to collect from the enterprise data sources – Microsoft and Slack – along with some other potential data sources like Box, Gmail, iManage, and others.

COLLECTING DATA

"RelativityOne's Collect tool can plug into any necessary enterprise data source, so when you receive that court order for all documents you can be confident that you can get all of the enterprise data sources you need via a four step wizard, which seamlessly and securely captures the data from those enterprise data sources," said Colloff. Additionally, since Relativity's SaaS platform is hosted in Azure, the Microsoft sources never leave the Azure secure cloud.

"~~However, as~~ much as we'd like all of the documents in a case to reside in those enterprise data sources, there are always going to be other documents," he added. "Someone - the client, general counsel, or somebody from the in-house legal department - is likely going to email you 50 PDFs to be loaded into the workspace and reviewed. RelativityOne's express transfer web import workflow allows those documents to be loaded right through the web interface without installing any additional applications."

UPLOADING DATA

When uploading data into RelativityOne, users can choose the destination folder, assign a custodian, and select the PDFs from the PC or the network. Simply point-and-click and the data will automatically be uploaded, processed and ready for review along with the other enterprise data sources. All the data will be in one place – even though it came from different work streams – without installing other applications or performing additional coordination or orchestration work. Once the data has been processed and is ready for review, RelativityOne can manage the various review queues and levels of review common in a complex eDiscovery project.

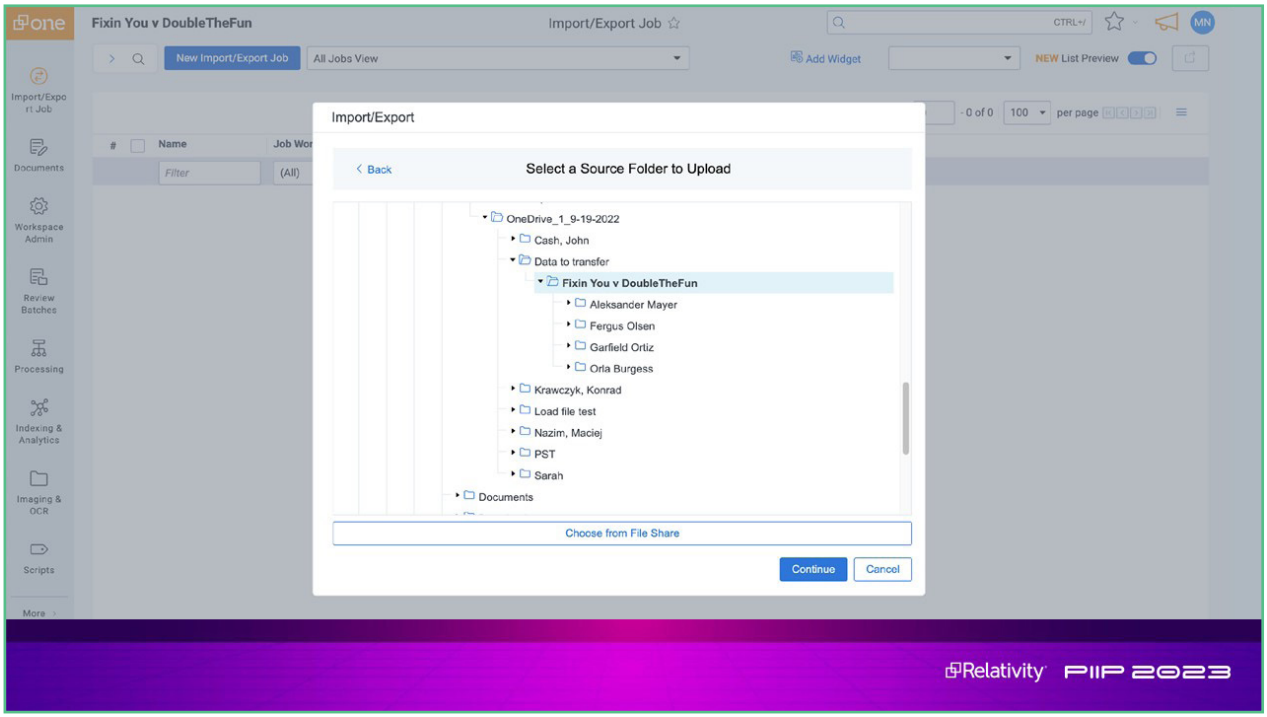


Figure 1.1 – The best eDiscovery technology makes project management more efficient, repeatable, and scalable.

In the investigative phase, when users are trying to find the documents that might have left the system when they shouldn't have, a new queue can be created in RelativityOne. According to Collhoff, "Since the hypothetical involves locating documents related to the theft of trade secrets, we can use the template 'IP theft search trade secret,' [see in *Figure 1.2*]– which might have been built for a previous matter or was created for us by CDS – that will combine sentiment analysis and other AI classifiers to find the documents that are most likely to be relevant to this kind of matter."

Once that queue is created, getting documents to review is simple – users just go to the review center and click on start review. Based on the AI and sentiment analysis, the most relevant documents will appear first. Review administrators see a detailed dashboard that shows the speed and progress of all the reviewers working on this queue of documents. This is helpful when determining high and low performers, deciding which reviewers to keep on (or not), and designing a more efficient way to manage the review.

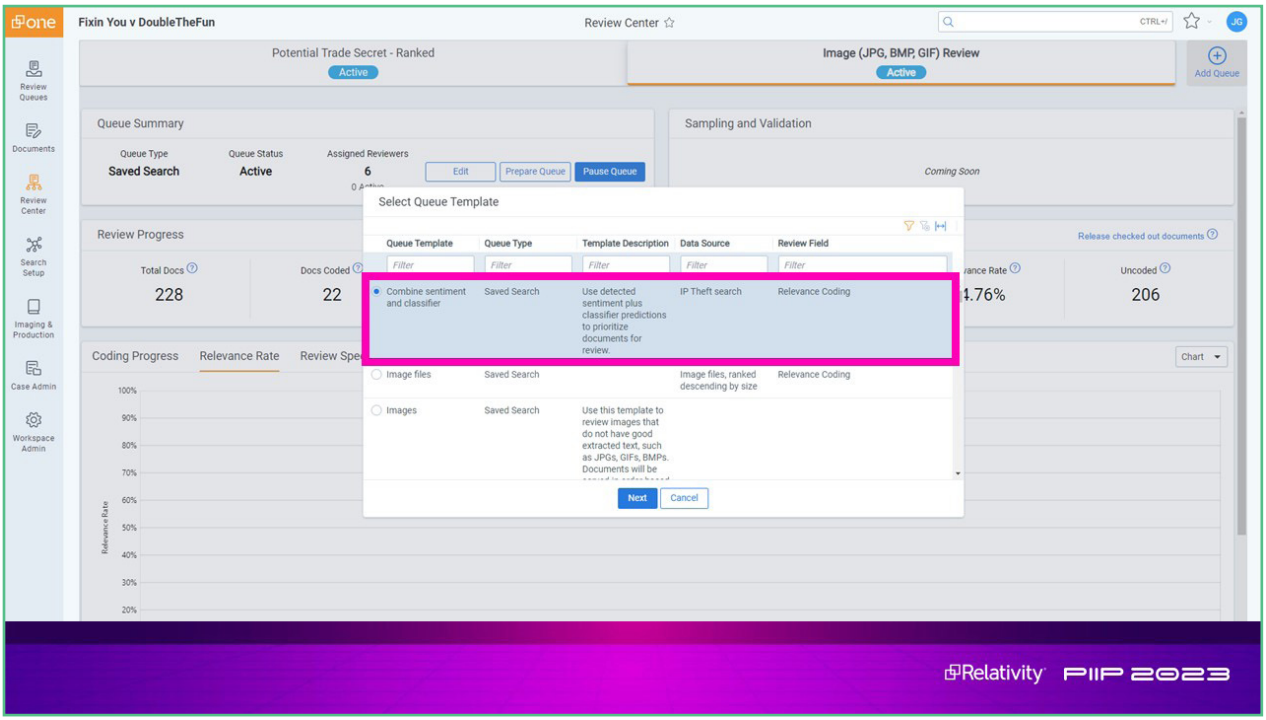


Figure – Relativity offers templated search configurations to apply and reapply as needed.

REVIEWING DATA

Many of the conversations in the world today are happening via some sort of a short message format like text messaging, Slack, Microsoft Teams chat, WhatsApp, and more. Said Collhoff: “Relativity developed the Relativity Short Message Format, the RSMF, to allow the industry to process and review short messages in a robust way. When we collected those Slack messages up at the top, they were converted by Collect to a short message format, so we can view them in our short message viewer here.”

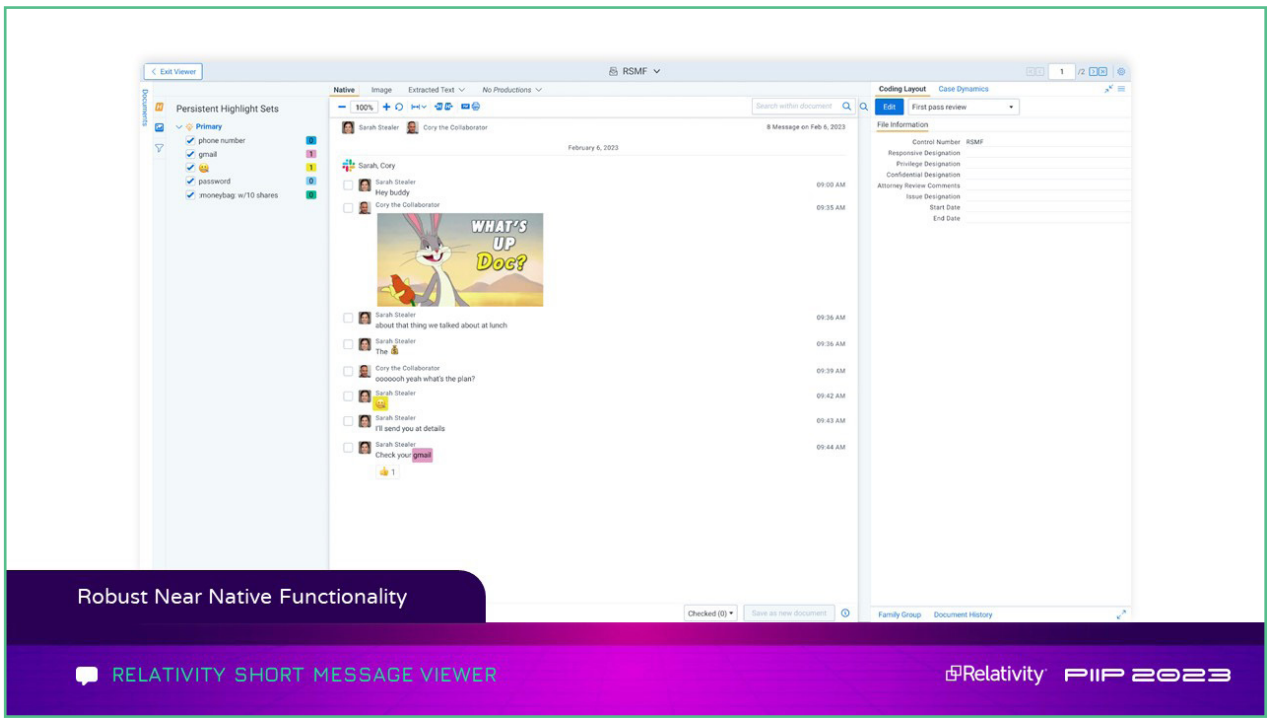


Figure – RSMF Viewer renders chat messages with emojis, reactions, and attachments inline.

“We see that Sarah the stealer and Corey the collaborator are chatting. We can keep track of who’s who based on their face avatars, we can see gifs and emojis and reactions just like you would in a native Slack conversation. So Sarah says, ‘Hey buddy,’ and Corey responds with a ‘What’s up doc?’ Giphy response. And Sarah says, ‘About that thing we talked about at lunch,’ the dollar sign emoji. And Corey goes, ‘Ooh yeah, what’s the plan?’ And Sarah says (with zipper face emoji, which is highlighted because we can search for and persistent highlight emojis as part of RelativityOne), ‘I’ll send you the details, check your Gmail.’ And Corey responds with a thumbs up emoji here.” If you didn’t have the context from this short message viewer, you might miss the important information.

“For example, in a text only format with the images and reactions as attachments, key information would be missed, and it would be easy for someone plowing through documents during review to not flag how important this section is to the entire matter,” Collhoff said. “Also, this is likely part of a much longer thread in real life where this little snippet is part of 250 messages that were sent that day between Corey and Sarah. It’s easy to scroll through that, if you don’t have the context, and miss what’s important during this review.”

Some further digging can now be done to figure out, based on what has already been learned, what happens next. “Using our communication analysis, we can quickly find emails between Sarah and other email addresses and domains,” said Collhoff. “In this case, we’ve got Corey’s Gmail here on the right in green - Corey@gmail.com. So let’s drill into these messages.”

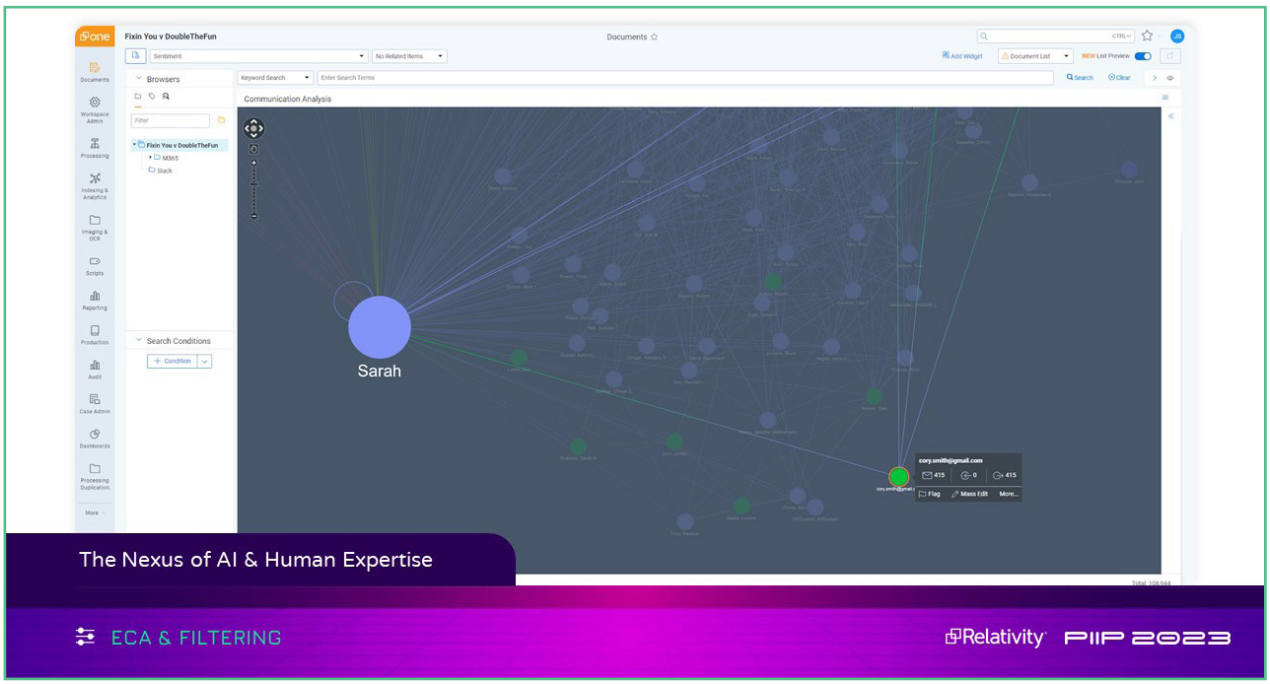


Figure 1: Communication analysis reveals who is talking to whom and what they’re talking about.

According to Collhoff, it is likely that the transmission of the transfusion technology document at the heart of this hypothetical will be found in that set of data. However, the email is sparse in content - with no subject, no text, and a details.zip attached. "Without RSMF context, it would be easy to miss, and trying to find it with concept search or sentiment analysis would be extremely challenging because there's nothing in this message, except the very important data that you're hoping wasn't stolen but in fact was stolen as an attachment," he said.

TRANSLATING DOCUMENTS

Since the hypothetical deals with data from all over the world and custodians who are multilingual, reviewers will likely encounter documents in various languages.

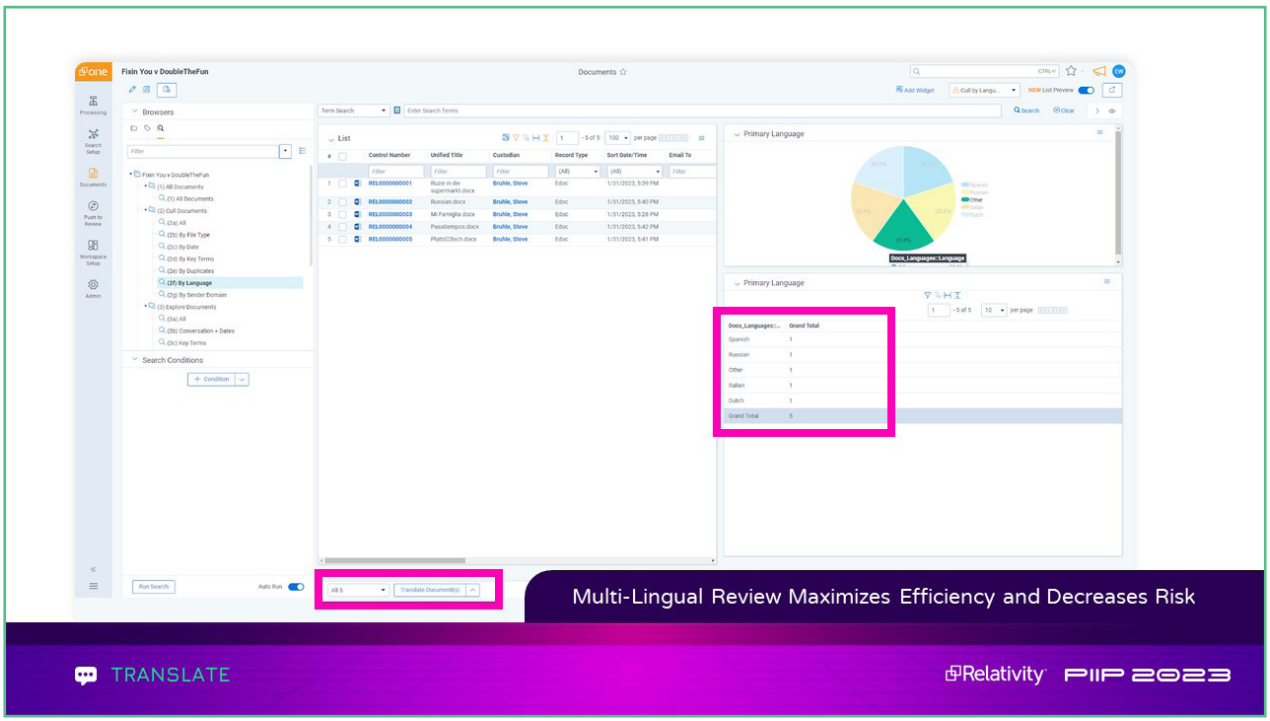


Figure . – Multi-language reviews can be managed efficiently using Relativity Translate.

Relativity Translate can help translate those documents. The right side of the slide above shows the different languages that were identified in the hypothetical dataset. "You can just bulk check them, translate all of them, and go through a translation workflow," said Collhoff.

However, because RelativityOne uses Azure's cognitive services behind the scenes, users will receive a native version of this document with translated text. For example, if you send out a PowerPoint that's in Italian, you get back a PowerPoint that's

in English with the same graphics, the same font, the same style and settings, except the language is English. “From a context perspective, that makes a world of difference when you’re trying to understand what’s happening vis-a-vis having, for example, just a text translation of the extracted text of a document,” he explained.

REDACTING PII

For Relativity, protecting the PII of customers and employees in these types of cases is paramount. RelativityOne’s automated redaction tool will find and redact social security numbers and other common PII, and can also create custom patterns to find and redact.

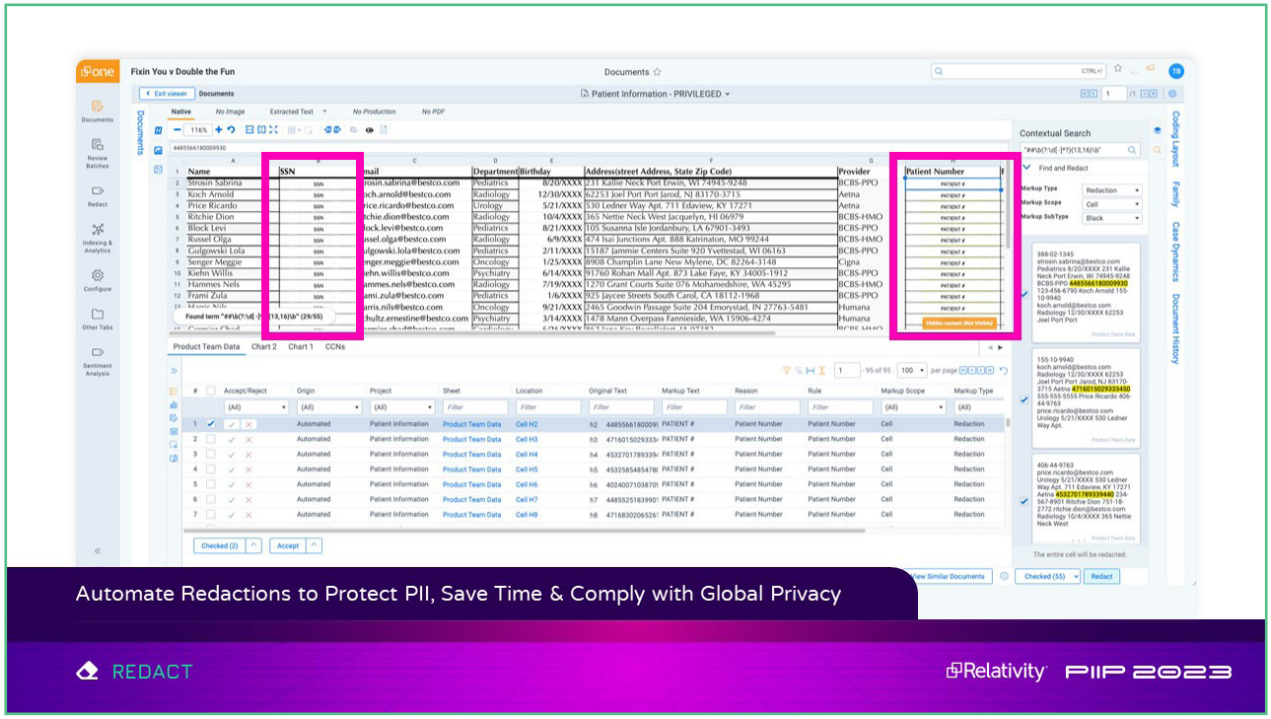


Figure 1.6 – To protect PII and comply with privacy regulations, deploy automated redaction.

“In this hypothetical, we’ve got a patient number also set up, which might have been something proprietary that FixIn You uses but other companies don’t,” said Colhoff. “But as long as we know what those numbers look like, RelativityOne or CDS can help you create a pattern to find all of those numbers and automatically redact them so you can protect all the PII, even if it’s not say the standard kind of name, date, social, et cetera.”

DISCO'S APPROACH

Kristin Zmrhal, vice president of product strategy at DISCO, discusses how the DISCO platform could help FixIn You identify relevant evidence.

COLLECTING DATA

First, according to Zmrhal, DISCO will create a core document team responsible for reviewing those documents to meet FixIn You's production obligations and leverage some of the prioritization to find key evidence as quickly as possible. The core document team will be fed by a data analytics team responsible for analyzing anomalies and trends in the documents, promoting documents into the actual review, and looking for things like the foreign languages and the documents that are identified as unknowns. "Triaging will be done as needed, with support from a forensics team responsible for the data collection and analysis of those forensic devices," she said.

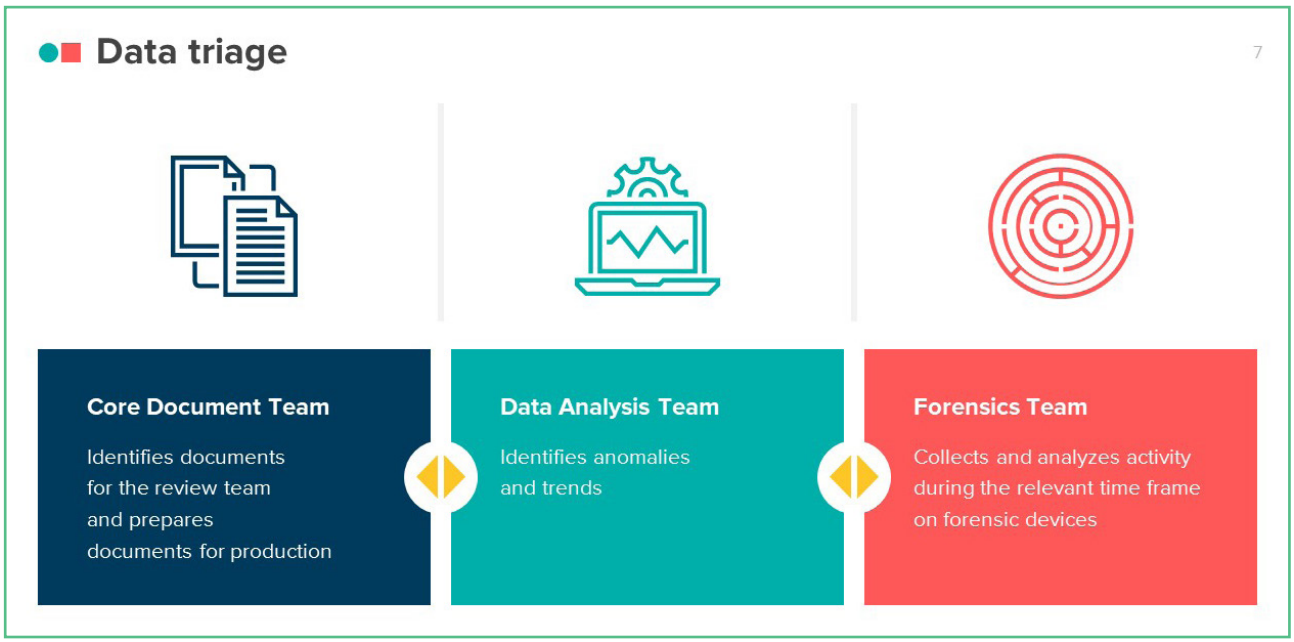


Figure 2.1 – In getting to know your data, DISCO advocates a team approach.

Next, the forensics team will collect physical devices and run analytics on them. You can directly upload cloud data sources into DISCO or local data storage. According to Zmrhal, "The forensics team is going to be looking for trends, deleted analysis, and what kind of devices were plugged in to see if that information needs to be taken to depositions and interviews with both the fixed and new custodians as well as the BloodFast team." DISCO will be available to provide deposition support or affidavits as needed.

HOLD NOTICES

The next order of business will be to ensure that the appropriate litigation hold notice has been sent out. The DISCO hold notice enables questionnaires to be forwarded to individual employees on the team related to the research and, based on what those questionnaires reveal, can be expanded to ensure all relevant data is preserved.

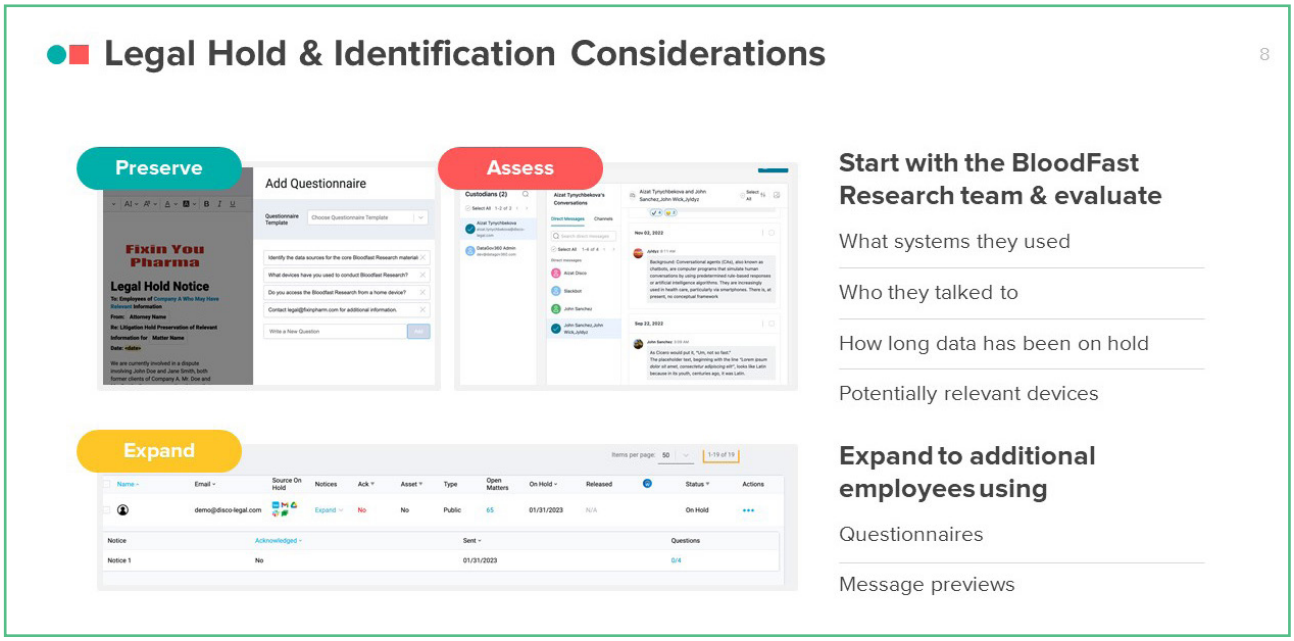


Figure 2.2 – Managing the legal hold process in DISCO

This functionality applies to data types from chat and collaboration tools as well. “The short message format from Slack and Microsoft Teams can be assessed during collection to get a quick peek into who was talking about what and what are likely the relevant channels to go and investigate once the data is moved into an eDiscovery application,” Zmrhal said.

UPLOADING DATA

The high-speed uploader will transfer the 16.5 terabytes directly into DISCO’s eDiscovery platform. During transfer and ingest, various steps happen automatically, including extracting containers, identifying metadata and text, de-duplicating data, identifying languages, and generating near native PDFs and a near native Excel viewer, which allows users to automatically redact for PII. A series of enrichments will also be run automatically on that 16.5 terabytes of data, including calculating the similarity within the dataset, threading email conversations together, and detecting hidden content, e.g., hidden sheets or speakers notes. DISCO will OCR everything, create those images, and build the entire analytics suite.

With about 16.5 million documents uploaded from all of these disparate data sources, after deduplication there are approximately 2.8 million documents in the DISCO ECA data space. Using DISCO's AI-empowered document review, the analytics team will proactively prioritize and continuously promote documents into the review population. "While approximately 770,000 documents will likely get promoted into review, by leveraging the prioritization and the AI methodologies, only about 130 of those will be reviewed by the DISCO managed review team, based on statistical analysis and historical data," said Zmrhal.

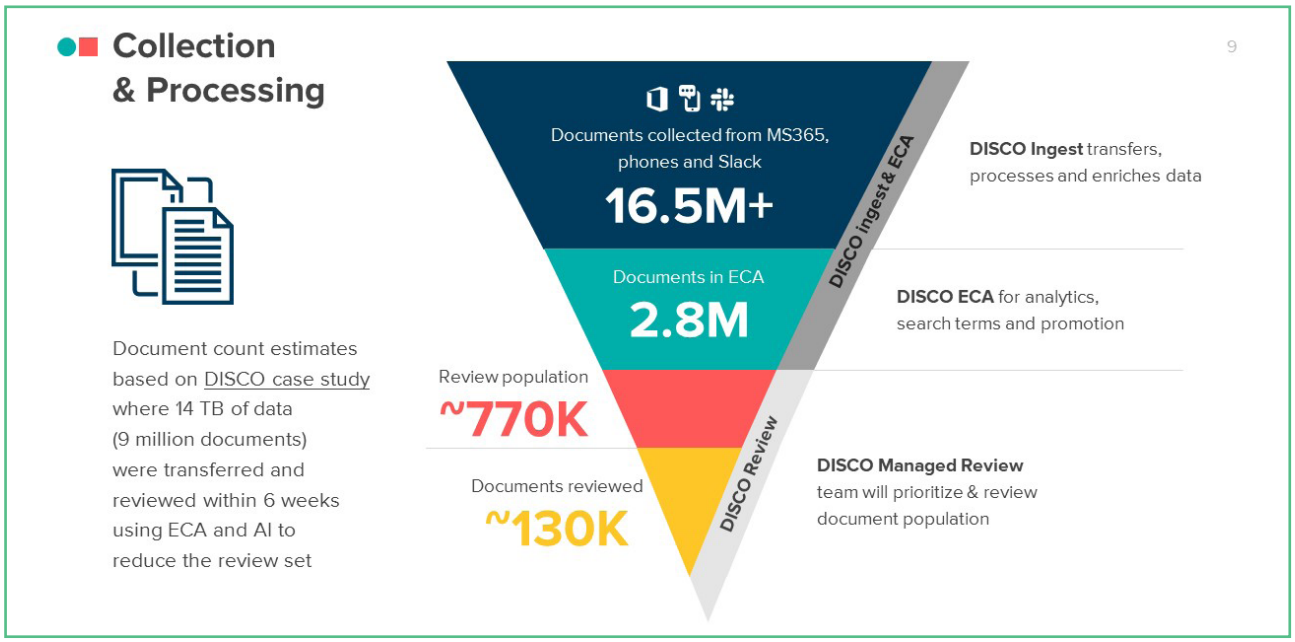


Figure 2.3 – DISCO used a recent case to evaluate what might occur in the hypothetical.

REVIEWING DATA

Prior to review, DISCO automates the application of analytics using integrated AI. "Whenever you upload data into DISCO, enrichments on the entire data set are run automatically, providing topic clustering with automatic indexing to dive into an automatic index or topic table of contents of that data set," Zmrhal said.

Documents are incrementally promoted by the data analysis team into the active review data space. "The promotion happens automatically once you hit those buttons," said Zmrhal. "It's all within the same database, two different data spaces. As you're finding pockets of data within the topic clustering or the enriched analytics, you're incrementally adding to that review workflow," she added. DISCO helps ensure that the relevant document categories are created by using issue tags and turning tag predictions and tag suggestions on before review begins. This allows users to prioritize documents within the relevant categories.

In the hypothetical, there are chemical design documents, research data, formula creation and specifications and facilities that are of particular interest to the client. According to Zmrhal, "As the reviewers are going through the documents, they will be tagging for those five different issues, and every batch they pull is going to have even a richer data set as it is learned from the activity of every single batch checked in. The tag suggestions are leveraged throughout DISCO's quality control processes, and the platform automatically flags discrepancies, anomalies, what the AI models are thinking, and what the humans are tagging," she said. "This allows reviewers to quickly identify potential issues in the document review and production."

SEARCH VISUALIZATION IN DISCO

Within DISCO, the search visualization page leverages rich analytics and data extracted during ingestion. As a result, the data analysis team can focus on the communications that occurred between the key custodians in their external email addresses or external parties. By clicking anywhere in search visualization, metadata widgets, the topic widgets, and the histogram widget can be layered on top. This will ensure that the documents being honed in on are likely relevant to the investigative of this case and fulfill document production obligations.

According to Zmrhal, reviewers will conduct a deep dive into the two key custodians who left and joined the other organization. "Their emails would be prioritized in a departed employee email set and put directly into my document review workflow. This is critical because in the hypothetical, they said that we did not agree on search terms, and robust analytics would be needed to identify those related documents," she added.

In this ~~particular~~ use case, reviewers would be looking for areas to ~~hone into~~ documents that are likely relevant ~~based on what they found~~, and also documents where the AI models and the humans do not agree to ensure that anomalies are found, errors are corrected, and the highest quality of review is ensured. Once the full review is complete, project managers work with the outside counsel to prepare a production set based on saved searches, which can be done in a rolling fashion throughout the life cycle of that review.

TOPIC CLUSTERING

The next step for the analytics team might be a closer look at topic clusters automatically created upon ingest. Topic clustering with DISCO's automatic indexing sorts the documents into topics and subtopics like this AI-generated table of contents:

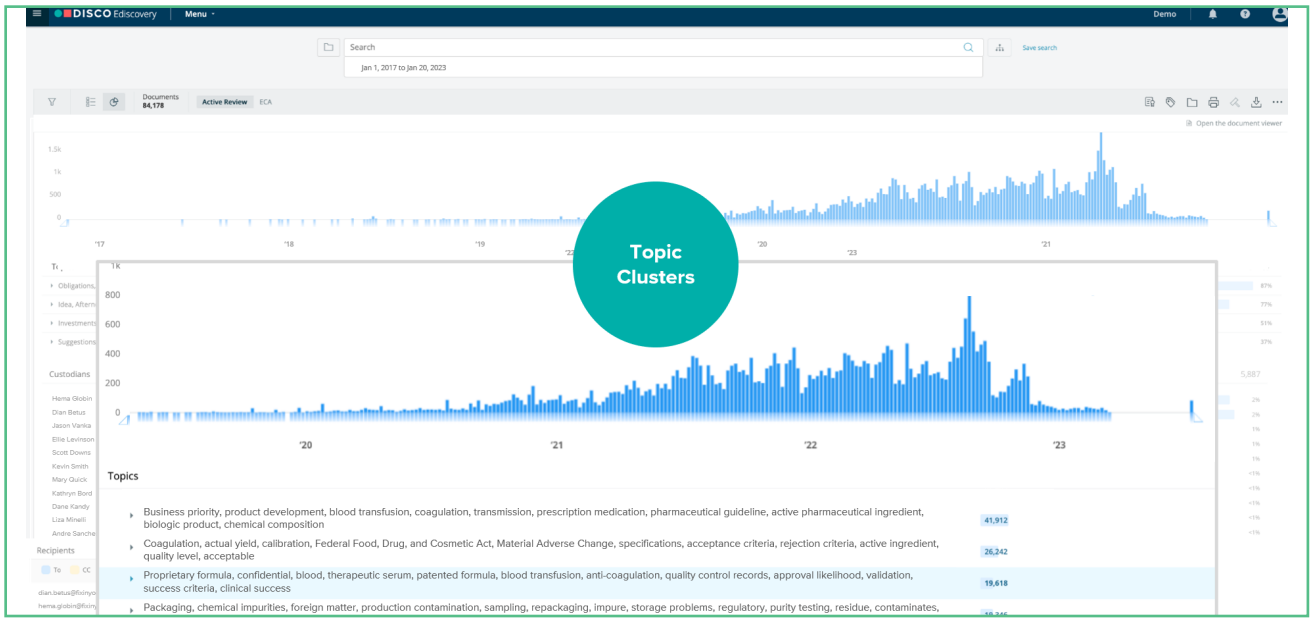


Figure 2.4

"The analysis team will use this to drill into likely relevant topics and promote them to review," said Zmrhal. "For example, you might see a topic that includes things like proprietary formula, anticoagulation, and clinical success, and if those topics seem important, you can one click promote to your active review space and put them into your review workflow."

Next, the analytics team will look at available sentiment analysis to determine if disgruntled employees might have sent notices within within email, Slack, or Teams before they left. Interactive filters applied through robust searches or directly through the visualizations help drill down further.

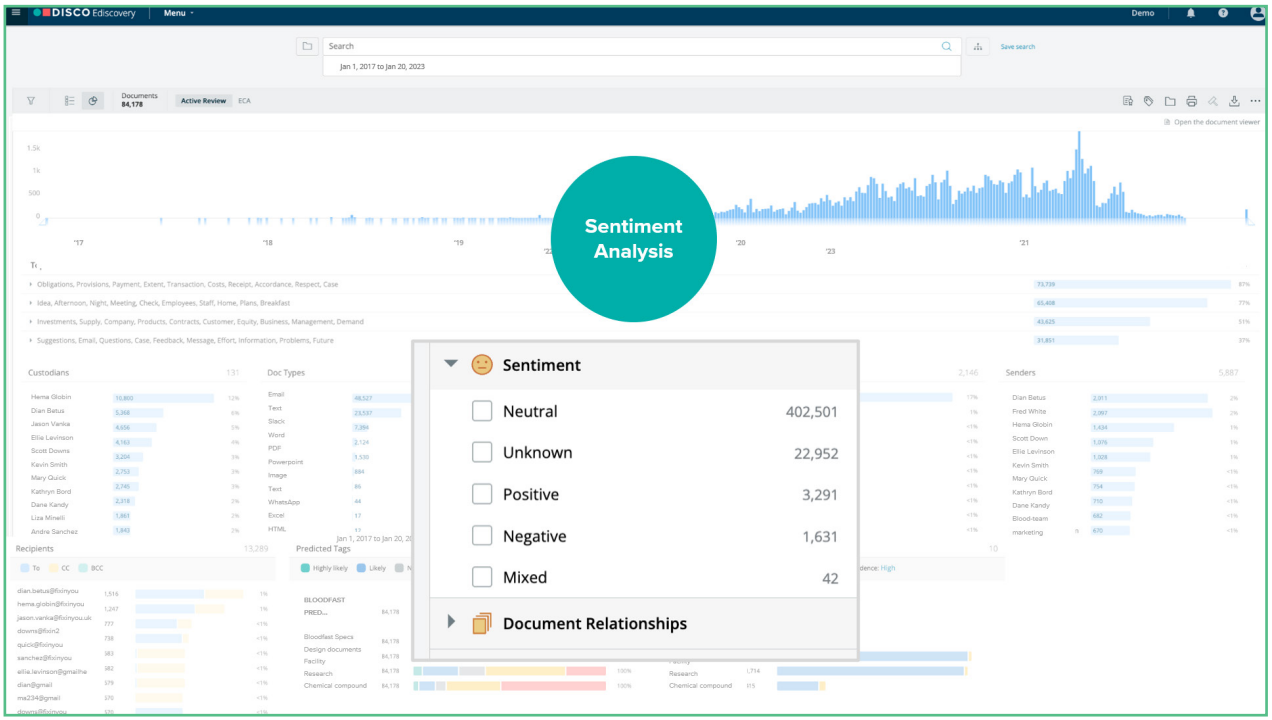


Figure 2 – As Sentiment analysis allows reviewers to quickly identify potentially responsive documents

TRANSLATING DOCUMENTS

Regarding potential languages, DISCO will automatically calculate the languages in the documents as they are ingested and the analytics team might triage which documents go to the review team and those that need further review. “For instance, the hypothetical mentioned Low Saxon as a language that might be available, “ said Zmrhal. “These would be looked at as undetermined, and if Low Saxon is not in the library, those would be triaged another way.”

As reviewers in the analytics team are going through documents, if they see a foreign language, a one-click button will translate directly into the document viewer so that a determination can be made whether the document is relevant to the issues. Zmrhal continued, “If you had an agreement to produce those documents in English, you might want human translation later. What we’re giving you here is a quick way to evaluate what the document means so that you can make those decisions in real time and as quickly as possible.”

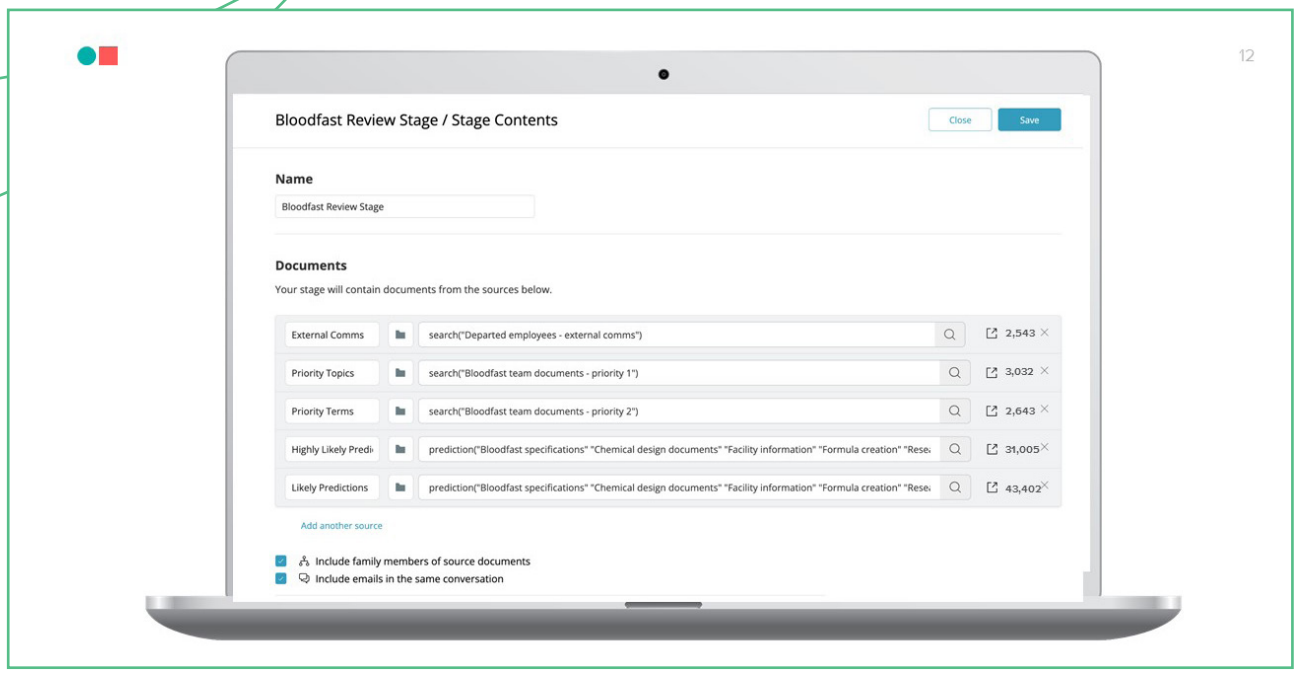


Figure 2 – An example of what building a review stage at DISCO might look like.

"Here, we've one-clicked promoted documents into the active review space and can now batch those out for our core document review team," she said. "The documents containing external communications from those key custodians that had left the company were prioritized, as were documents based on other search terms or different topic clusters that were potentially relevant. The last two lines display plain English feeding into the review team documents that the machine has predicted as highly likely or likely to the previous five issue codes."

Every document in DISCO will be scored on a negative-100-to-100 range – positive numbers mean it's more likely related to a topic and the negative numbers mean that it's less likely related to a topic. When these are added to the review stages, the likely relevant materials are brought to the top of that review workflow. "This saves time because as attorneys move through the documents, we'll see a sharp decline when we get to sort of the final pool of the relevant materials," said Zmrhal. "Through various tactics like prioritizing relevant materials and only reviewing the last in time messages, 770,000 documents potentially relevant based on search terms or topic clusters can turn into 130,000 documents that we actually need to collect. Because as attorneys move through the documents, we'll see a sharp decline when we get to sort of the final pool of the relevant materials," said Zmrhal. "Through various tactics like prioritizing relevant materials and only reviewing the last in time messages, 770,000 documents potentially relevant based on search terms or topic clusters can turn into 130,000 documents that we actually need to collect."

REDACTING PII

Because of the nature of the organization in the hypothetical, software companies need to comply with HIPAA and GDPR rules. As documents are ingested, a secondary index of common patterns like social security numbers and PII and PHI is being created. However, there is no need for individual reviewers to manually redact this information in DISCO. Instead, a mass redaction is applied, based on the common patterns.

Depending on how concerned the client is about the transfer of data, DISCO could also host this in one of its data centers outside of the United States and staff it with the company's global professional services team. Things like tag sampling would be used to run QC protocols, propagate tags by hash, and apply AI to ensure the highest quality production.

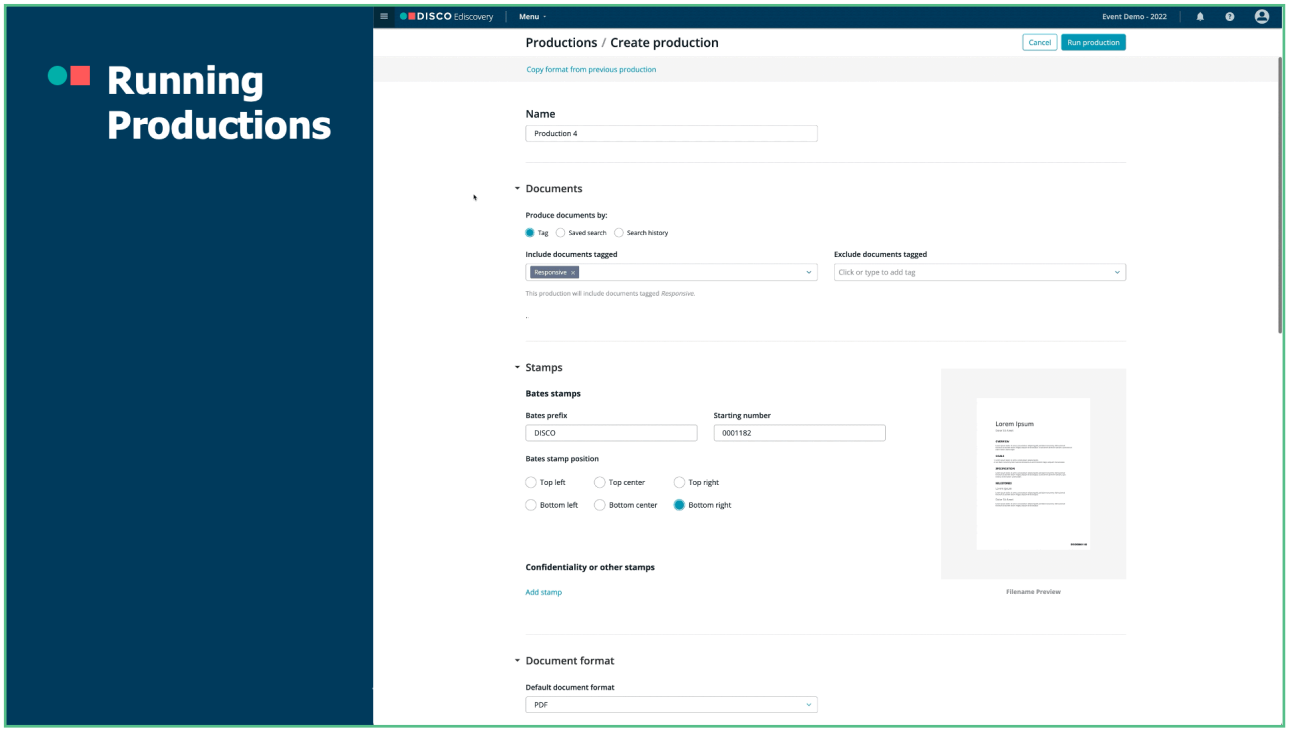


Figure 2.7 – According to Zmrhal, 85% of DISCO's productions take less than 30 minutes no matter how big or how small they are.

One final note: When DISCO receives the DoubleTheFun production, all of that data will flow in directly via native uploaders. DISCO also has cloud connectors as well as the ability to ingest productions from another side, and all of that will get populated into an analytics dashboard, as shown in the above image, so that multiple different areas can be scrutinized quickly.

APPENDIX

BIOS

Shamir Colloff, Senior Product Manager
Relativity

Shamir Colloff joined Relativity in 2017 and is a Senior Manager in Product Strategy, focused on RelativityOne, where he works closely with the Relativity user community to assess needs and business opportunities and help drive future product development priorities. Prior to joining Relativity, Shamir was the chief technology officer for a Relativity Hosting Partner, Evolve Discovery. He has extensive large-case management experience with Am Law 100 law firms and Fortune 500 General Counsels, and has supervised multi-terabyte projects in all phases, including data collection, processing, analysis and production. Prior to joining Evolve Discovery, Shamir was at Morrison Foerster, where he provided litigation support and tech consulting to their Silicon Valley office in intellectual property and internal investigations.

Kristin Zmrhal, VP of Product Strategy
DISCO

Kristin Zmrhal brings over sixteen years of experience in legal technology consulting and support and a passion for leveraging technology and optimizing workflows. She spent seven years at Google and led Google's litigation support operations practice, where she built a team of project managers and technical specialists who used technology to reduce Google's litigation spend while improving efficiency, quality, and data security. Prior to Google, Kristin spent seven years as a consultant for a number of Fortune 500 companies and Am Law 200 firms.

THE HYPOTHETICAL, EXPANDED

FixIn You Pharmaceuticals was on the brink of launching their first dynamic blood transfusion product, Blood Fast™ when DoubleTheFun Pharma launched a competitive product.

FixIn You and DoubleTheFun never interacted in any business meeting or anywhere other than on conference floors to be polite. FixIn You has determined that two former employees who worked on the product must have taken the research data and/or the formula when they left the company 13 months ago to work for DoubleTheFun. FixIn You, to gain back control of Blood Fast sues DoubleTheFun. After surviving a Motion to Dismiss, FixIn You must now examine the Blood Fast research team's documents (including the two former employees) to determine when and how their data was taken out of their facility/ies.

A total of 17 custodians' data must be reviewed from custodians residing in the United States, Italy, and the Netherlands. Data sources are MS365 (Exchange, Teams, Excel, Word, and SharePoint), and company issued mobile devices (iPhone 13 and iPad) and Slack data will cover the data universe for this suit. Search terms cannot be agreed upon; counsel should be very wary of how they examine data looking for shorthand references to any Blood Fast information and/or the actors in the alleged theft. Some of the FixIn You research team members live outside the United States and may have communicated in Italian, Spanish, Dutch, Low Saxon and Russian. Total data size at collection: 1.65TB.

MATTER REQUIREMENTS

FixIn You counsel needs to determine their approach. Some considerations for the legal team include:

- Process the data for all 17 custodians
- Analyze any communications with external parties
- Identify outlier emails to unknown domains or to anonymous web-based email addresses
- Determine the source of the data leak
- Determine, based on the data sets expected from the defendants, what data was provided and confirm the source of the leak. (Once identified, FixIn You will modify the complaint, adding the name/s of those who shared proprietary information).
- Identify any emails from DoubleTheFun where they knew that the product being developed by their employees was being done with stolen, proprietary information
- Identify any additional issues or information for other products which may have been shared to protect their proprietary information
- Leverage analytics tool/s to gain faster access to the most important data
- Eliminate product team members still employed from the list of potential sources of the leak
- Produce data on a rolling basis
- Pursue an aggressive 4-month timeline for production and completion of document discovery
- Provide a QC protocol leveraging as much information as possible and protecting patient information
- Maintain the HIPAA requirements in produced data
- Comport with GDPR and identified countries data privacy statutes

ASSIGNMENT

Tell us how your platform addresses this data and provides unique insights into the data sources and or set. Specifically, be prepared to address:

1. The estimated time for ingestion and processing, estimate review time with a data set of 83% similarity.
2. How much time is needed to de-duplicate, de-NIST and make the 1.65TB data set review-ready?
3. The quickest, most accurate method for approaching review, including when and how you may deploy AI and/or analytic technologies before reviewers begin their work, and then how those technologies can assist in that work.
4. Redaction, especially related to HIPAA and privacy protected data.
5. How your tool addresses text messages and/or chat data from the data set, and how you handle images, avatars, or video.
6. Identify a way to identify documents which contain Blood Fast's chemical composition.
7. How you will handle the foreign language data.
8. How you will identify the foreign language documents by language type.

Note: All the tools leveraged must be a part of your platform and not third-party applications or tools.

ABOUT PIIP

CDS' signature annual event, The Putting Insights Into Practice Forum, also known as PIIP, focuses on the business, legal and data challenges that converge in eDiscovery. Industry thought leaders offer technology-forward strategies that can help practitioners in law firms, corporations and government work smarter and elevate their eDiscovery.

www.cdslegal.com