Practical Advice for Non-English Language Electronic Discovery Projects

This white paper discusses:
- *Collection and transfer of data outside of the U.S.*
- *Search design and translation challenges*
- *Translation and document review*

in this paper

INTRODUCTION

If you are a U.S. based IT or litigation support professional, you have likely received an email message that ended with "…and there are three people in the Beijing office." If you have not, you probably will.

Non-English language electronic discovery continues to be on the upswing, driven by several business challenges including:

- Increasing international IP litigation, most prominently reported upon in the technology industries such as cell phones;
- Ongoing cross-border M&A activity;
- Increasing enforcement of FCPA cases by the U.S. and the U.K.

COLLECTION

From a technical perspective, there is little difference between collecting electronic documents within the U.S. and internationally. Products such as EnCase® can be used to create either a physical (bit-by-bit) or logical (file-by-file) collection of a hard drive. The primary differences are typically cultural.

In some EU countries, a representative from the client's HR department, the local Works Council, the government or all three may be present while data collection is performed. In many EU countries, the collection schedule must be planned for normal business hours. Holidays, evening and weekend access typically must be scheduled far in advance and may not even be allowed depending on the country. This reality limits the time window for data collection to about six useful hours per day.

Due to the cultural nuisances, using a local forensic collection expert, if available, is often the most effective method to collection outside of the U.S. A key to improving collection efficiency begins early in the planning process. A project manager from the legal team must work with the client to identify all IT staff who may be necessary to support a collection effort.

EUROPEAN UNION PRIVACY

In some situations, the court in the local EU country will only allow potentially responsive, relevant documents to be transferred to the U.S. The court may require that processing and document review be conducted either within the country or at least within the EU. This may also be part of a consent agreement, whereby the employee only agrees to the collection as long as those items agreed upon as "private" are removed prior to transfer to the U.S. This situation is typically managed by use of a hosted solution using reviewers in the EU or a war room setup within the country. When first pass review is being performed onsite, it is necessary to have an attorney present who has the authority to make immediate decisions regarding document coding. To improve the defensibility of the onsite activities, you should create complete and precise documentation of how onsite processing, searching and review were performed.

BRINGING DATA BACK TO THE U.S.

Unless the amount of data is very small, the data will need to be either physically shipped via a reliable carrier or personally couriered. While not quite as safe as being carried by a courier, a shipping carrier can often be used to physically ship data to the U.S. While uncommon, packages can be lost or delayed for short periods of time in U.S. Customs.
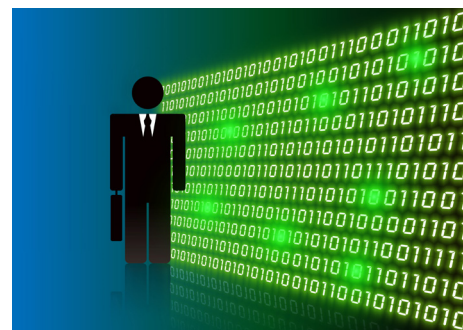
If time is not on your side, then a courier must fly to the U.S. In a pre-9/11 world, the primary concern of physically transporting data collected in Europe was the risk of damaged media. The new additional risk is one of seizure. One of the many tools provided to the Department of Homeland Security to combat terrorism was the power by the U.S. Immigration and Customs Enforcement division to seize and search computers and documents transported into the U.S. A number of recent news stories have highlighted how travelers into the U.S. are having their computers seized.

There are two primary steps to follow to reduce transportation risk. First, ensure that at least one additional set of media resides within the country at either the office of your firm, local counsel or vendor. Second, ensure that the media, either shipped or couriered by a

person is accompanied by documentation from the law firm.

UNICODE PROCESSING

One of the issues with non-English language electronic discovery is how non-English language characters are interpreted by the processing and review applications. Computer terms often have a magical ability to glaze over the eyes of a legal professional. The three terms below are important to understand as they are methods by which computers translate human language into computer language.

Prior to Unicode, data was stored and displayed using either ASCII or ANSI. ASCII was used on the first IBM PC's as a standard for English character sets. ASCII could be extended to support non-English character sets, but there was no standard for doing so. This made it virtually impossible to send electronic documents internationally. ANSI later provided standards for extending the ASCII character set to support non-English characters. By installing the appropriate language pack on a computer, ANSI made international electronic document transfer possible.

Unicode is an open standard managed by the Unicode Consortium, a non-profit organization that is attempting to provide a universal numeric representation for every character of every language on the planet. For example, Unicode allows a person using English as their default computer language to receive an email written in German and the German language characters will properly display.

Here are a few obscure, and not so obscure, items to know about Unicode:

- Microsoft Windows NT, Windows XP, Vista and Windows 7 support Unicode.
- Microsoft Windows 95, 98 and ME had limited Unicode support.
- Microsoft Office 97 - Microsoft Word, Excel and PowerPoint supported Unicode, Access and Outlook did not.
- Microsoft Office 2002/XP added Unicode support for Access and within message bodies in Outlook 2002.
- Microsoft 2003 provided full Unicode support in Outlook 2003 as long as the user was on a Microsoft Exchange 2000 email server.

The technology used for processing non-English electronic documents must be able to interpret and correctly extract the text and metadata documents created in ASCII, ANSI and Unicode. This data must be stored in a Unicode compliant database. The indexing and query engine must be able to support Unicode in order to execute searches. Finally, the review application must be able to store and display Unicode characters.

Due to the nature of international business, it is common to find multiple languages within a single email message due to message threads. Because of this, the review team needs to be able to identify documents by primary language type in order to assign documents to the correct reviewer. Most technologies that identify languages within documents use one of the following methods. For speed purposes, the technology will scan a limited number of lines in the document and assign a language code to the document. For improved accuracy, the technology will identify a dominant or primary language based on the language found within a given document that has the greatest percentage of use. Ideally, the technology should also be able to identify and store in a searchable field all languages found within a given document.

SEARCH DESIGN

Diligence must be applied when designing and executing keyword searches. Besides the need for vetting by someone fluent in the language, there are a number of items to remember that can make a significant difference in accuracy.

Using native language speakers and/or professional linguists may be beneficial when designing slang terms for certain concepts, such as "kickbacks" and "bribes." This expertise is very valuable for languages such as Chinese, where regional dialects are common.

For basic keywords, like in English, alternative spellings should also be tested. For example in German, test terms with "ss" in addition to "ß" or "ue" in addition to "ü." In languages like French, test alternates to characters such as "à" or "è" by using "a" or "e." Due to spelling differences, British English keywords must be designed with equal diligence as German or French by American English speakers. A British English document might state "I recognise the theatre by the harbour" while the same sentence in American English would be written as "I recognize the theater by the harbor."

It can be confusing how dates and times are handled by a computer. October 12, 1954 can be displayed on an U.S. English computer as 10/12/54 but can be displayed on a Spanish computer in the European format of 12/10/54. Executing date and time filtering within metadata, such as the sent date and time on an email, is fairly simple as all date and time metadata should be standardized by the processing technology into GMT (Greenwich Mean Time) format. However, searching for dates within the body of a document may require that both date formats be used to ensure documents are not missed.

Designing searches for currency values sometimes require a great deal of patience and iterative testing. Some currency identifiers are simple to find in searches, such as USD, $, GBP, £, EUR or €. However, French francs are typically identified with the letter "F" as in 131,51 F and the Danish kroners are identified with the letters "kr" as in kr 131,51. Searching for the characters of "F" and "kr" will yield an unusable number of false search results. (The French franc was replaced by the Euro in 2002.) Also, did you notice in the examples that the currency values contained a comma? Most currencies use the same decimal and thousands separator that the numbers in the country use, but this is not always true. In some places in Switzerland, they use the period as a decimal separator for Swiss francs (Sfr. 131.51), but then use commas as the decimal separator everywhere else (131,51).

Searching for specific numeric values is slightly less complicated than currency, but still required fortitude. For example, consider the simple thousands separator. In the U.S., this character is a comma (,). In Germany, it is a period (.). Thus one thousand and one is displayed as 1,001 in the U.S. and 1.001 in Germany. In Sweden, the thousands separator is a space and 1 001 would be displayed. The decimal character is equally as interesting. In the U.S., this character is a period (.). In Germany, it is a comma (,). Thus one thousand one and eight tenths is displayed as 1,001.8 in the U.S. and 1.001,8 in Germany.

TRANSLATION

Machine language translation is often useful to assist English-only speakers with the substantive issue coding by the U.S. legal team. A limited number of vendors have the ability to perform machine language translation of acceptable quality that meets the needs of document review. An even smaller number of vendors have document review applications that allow the reviewer to submit a block of text for just-in-time translation. Regardless of the approach used, machine language translation is a literal translation with accuracy similar to the early days of OCR. The benefit of using machine language translation is that the number of exceptions that must be reviewed can be significantly reduced. Regardless of which of these methods the legal team uses, have a sample of data from your matter processed by a vendor and evaluated for acceptability by the review team prior to committing to any one solution.

Documents that will be presented in depositions or as exhibits in court should be translated by translators certified by the American Translators Association who are also court-certified. This provides the highest quality of translation; ensuring language nuances such as idioms are translated as they are used instead of relying on a clumsy and inaccurate literal translation.

DOCUMENT REVIEW

Attorneys who are native speakers of Asian, Middle Eastern and Russian languages have been very busy over the past several years. Competent document reviewers with these particular language skills continue to be at a price premium. Native speakers are more capable of understanding idiomatic expressions, slang and possibly the nuisances of local dialects when reviewing email communications and other types of informal correspondence. For larger reviews, it may be more practical to use machine language translation for first pass review and use native speaker reviewers for escalation, QC and second pass review.

CONCLUSION

This paper is only a brief summary of non-English electronic discovery. In any matter involving international electronic discovery, legal teams should retain local council. If the U.S. legal team does not have experience with international electronic discovery, a consultant or vendor with expertise and references in this arena may

be critical to executing a defensible and efficient discovery project.

The Author

Alan Brooks
Vice President, Marketing

As Vice President of Marketing, Mr. Brooks is responsible for developing and managing integrated marketing and training programs designed to educate clients on complex data management challenges and solutions, including Early Information Assessment℠ methodologies and managed services solutions. Under his marketing leadership, Mr. Brooks utilizes his expertise to further enhance the company's position as trusted advisor and innovative strategic partner to corporate and law firm clients. An industry veteran, Mr. Brooks has 20 years of experience in marketing, sales and product management in the high-tech and legal industries providing thought leadership, and new marketing and educational initiatives.

Mr. Brooks has a proven track record identifying business growth opportunities for companies. He has also successfully developed lead generation programs integrating Web marketing, email campaigns, telemarketing, Webinars, seminars, executive briefings and workshops. Mr. Brooks also has hands-on domain experience with processes and technology for all areas of the EDRM model from Information Management through Production. His experience includes SEC 17a-4 compliant archiving systems, information governance, data extraction and migration and designing proprietary electronic discovery applications.

Prior to joining eTERA Consulting, Mr. Brooks was the Vice President of Marketing at Sfile Technology Corporation and held similar positions at RenewData and CaseCentral. While at RenewData, Mr. Brooks was a member of the management team that grew the company from $5 million to $30 million in revenue in three years. While at CaseCentral, Mr. Brooks successfully repositioned the company and helped them grow from a non-profitable $5.2 million to a profitable $12 million. Mr. Brooks has also held executive and management positions at Remedy Corporation, Utopia Technology Partners and Pfizer.

About eTERA Consulting

eTERA Consulting specializes in helping organizations improve information governance, compliance and discovery management. We are a technology independent consultancy with a broad range of services from strategic information consulting to project-based engagements. Because we are not locked into a particular vendor's technology, our clients benefit from flexible service delivery and pricing options.

A unique differentiator for eTERA Consulting is our Early Information Assessment℠ (EIA) methodology which we use to help clients implement a proactive approach to integrating the management of corporate information as it relates to risk management, regulatory compliance, electronic discovery, litigation hold, records management and IT storage.

Headquartered in Washington, DC, eTERA has served the legal industry since 2004. The company also maintains additional offices across the United States. eTERA was named to The Inc. 500 list of fastest-growing private U.S. companies in 2010.

DISCLAIMER: Legal Information Is Not Legal Advice

eTERA Consulting is not a law firm. This bulletin provides information to help professionals better understand the intersection between technology and legal processes. But legal information is not the same as legal advice -the application of law to an individual's specific circumstances. Although we go to great lengths to make sure our information is accurate and useful, we recommend you consult a lawyer if you want professional assurance that our information, and your interpretation of it, is appropriate to your particular situation.

eTERA Consulting
1100 17th Street, NW
Suite 605
Washington, DC 20036
Direct: 315-566-9330
info@eteraconsulting.com
www.eteraconsulting.com