Multilingual Ediscovery: Options, Obstacles and Opportunities Report

A guide to collecting, filtering, reviewing and producing multilingual documents in discovery.
This document does not provide legal or other professional advice and should not be relied upon as anything other than a starting point for research and information.

No part of this publication may be reproduced, transmitted, transcribed, stored in a retrieval system or translated into a language or computer language, in any form by any means, electronic, mechanical, optical, chemical, manual or otherwise, without the express written consent of Kroll Ontrack Inc.
Introduction

When Magistrate Judge Kiyo Matsumoto of the Eastern District of New York compelled a French bank to produce documents in 2007 to victims of a terrorist attack in Israel, the issue of the global courtroom became quite clear. [See Strauss v. Credit Lyonnais SA, 242 F.R.D. 199 (E.D.N.Y. 2007).] Lawyers are now responsible for litigation without borders and discovery without language barriers. The era of globalization is colliding with the complexity of ediscovery.

Nowhere is the language obstacle more pervasive than in the legal profession—and more specifically in the area of discovery—where documents created in any country and any language could be relevant to a lawsuit, investigation or regulatory matter. Legal teams must tackle this hurdle head-on in order to meet discovery obligations and provide the most effective representation for the client. As a result, attorneys and litigation support professionals are finding it necessary to become experts in everything from data privacy issues to the technology challenges faced when processing data in various languages.

The legal community is at a tipping point when it comes to discovery. Over the last few years, the technology to collect, filter, process, review and produce multilingual data has evolved quickly. However, many practitioners find themselves struggling to understand the revised Federal Rules of Civil Procedure, let alone the technical vernacular and legal implications associated with cross-border discovery. Unfortunately, globalization will not wait for the legal profession to advance its knowledge and develop best practices. It is only a matter of time before a case on your docket, in your firm or on your desk involves multilingual discovery.

This white paper will address current practices, define standard terminology, explain the available technology and present relevant considerations during each step of a multilingual discovery project: collection, filtering, processing, review and production. Armed with this information, you will be empowered to provide options to your clients, avoid obstacles and leverage opportunities when coordinating and managing discovery across the globe.
What Terminology Is Important to Know?

There is no better place to begin the technical discussion of multilingual discovery than with a vocabulary primer. Unicode, double-byte characters, code pages—what do these words mean and how do they impact an international discovery project? Why are so many different groups of legal and technical professionals using different definitions for the same words?

Unfortunately, when it comes to multilingual discovery, the industry lacks a uniform vernacular. The common law, regulatory requirements and litigation support best practices for handling multilingual data in discovery are still developing. In addition, the technology behind handling data in multiple languages is extremely complex. All of this results in a high degree of inconsistency and distress when it comes to these aspects of discovery. As articulated in a post on a litigation support chat group:

“Here are my thoughts on Unicode. This is an area where not enough knowledge and education is going on to inform law firms and even some service providers of what the heck Unicode really means.”

Therefore, it is clear that the first step in understanding multilingual discovery is to establish common vernacular by defining key words and phrases.

General Terminology

- **Multilingual Discovery**: When two or more languages are involved in discoverable documents; this is the industry-preferred term (also referred to as LOTE—“Languages Other Than English”).

- **Foreign Language Discovery**: When discoverable documents contain languages other than a speaker's native language; not the industry-preferred term.

- **EFIGS**: English, French, Italian, German and Spanish; common languages in the Western European language set.

- **Cyrillic Languages**: Russian, Bulgarian, Serbian, Turkish, Polish; common languages in the Eastern European language set.

- **Bidirectional Languages**: Some languages, such as Arabic and Hebrew, use a right-to-left (RTL) format, in which the writing begins at the right side of a page and concludes at the left side. This is different from the left-to-right (LTR) format used in English and most Western European languages. When LTR text is mixed with RTL in the same sentence or paragraph, it is known as bidirectional text. Without special programming, computer systems typically do not correctly display, recognize or search bidirectional text.

For more terminology, turn to the Appendix contained on page 14.
How Is Language Data Stored?

It is important to understand that computers hold text in ways that are completely foreign to their human users. It is the translation between how text is stored within the computer and the presentation of that information to the human user that enables a computer to store and display a variety of languages. There are two main methods most computers use in the successful handling of multilingual data—code pages and Unicode.

**Code Pages**

Code pages were one of the first methods used by computers for supporting different languages within a set of documents. Code page is a generic term used to describe a cross-reference table that translates the data within a file into human-friendly text. Code pages are numbered for reference, are used in most computer operating systems (e.g., Microsoft® Windows®, Apple® Macintosh® and Linux®) and commonly support the representation of a specific set of characters and/or languages. Each code page converts data to text using the same underlying data values, but generates different language characters based on the code page itself. For example, “code page 1252” on a Windows computer refers to the characters usually found in the Latin alphabet along with most of the accented characters in Western European languages, common punctuation and special characters. There is no universal set of code pages used by all computer systems, and as a result, files created on one machine may be unreadable on another. It is important that a wide variety of common code pages be supported for adequate multilingual support.

**Unicode**

Unicode is a standard agreed upon by computer hardware and software makers and other international organizations, where each character in nearly every possible language in the world is assigned a number. Unicode has become an industry standard, allowing computers to consistently represent and manipulate text expressed in the world’s writing systems. Instead of multiple code pages for multiple languages, the Unicode standard essentially acts as one large code page for all possible languages. Today, the Unicode character set consists of a repertoire of about 100,000 unique characters. Unicode resolves many of the issues relating to the simplistic code pages in that it handles much more than just characters and is more extensible for future growth.

**Explaining Code Pages and Unicode**

Electronic documents are generally saved in either a code page or Unicode format and this information is critical for accurate ediscovery processing. Below are sample languages and corresponding character set information.

<table>
<thead>
<tr>
<th>Microsoft Windows Code Pages</th>
<th>1252</th>
<th>1253</th>
<th>1251</th>
<th>1255</th>
<th>1256</th>
<th>932, 936, 949, 950</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM® Code Pages</td>
<td>850</td>
<td>737</td>
<td>855, 866</td>
<td>856</td>
<td>864</td>
<td>833, 836</td>
</tr>
<tr>
<td>Supported Languages</td>
<td>Latin, Western European (English, French, Italian, German, Spanish)</td>
<td>Greek</td>
<td>Cyrillic (Russian)</td>
<td>Hebrew</td>
<td>Arabic</td>
<td>East Asian (Chinese, Japanese, Korean)</td>
</tr>
<tr>
<td>Unicode</td>
<td>Unicode crosses all language boundaries and supports hundreds of languages.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How Is Language Data Displayed?

In addition to the basic encoding of different characters, whether by code page or Unicode, it is important to understand how those characters are presented for display. How individual characters are displayed is determined by font selection.

Fonts can be extremely important in multilingual discovery since font selections can include critical components such as underlining, bold and italics, as well as the stylistic representation of characters. While code pages and Unicode determine what characters will display in metadata and text extraction, neither character encoding method specifies what typeface is used to represent the characters. If a font does not include a particular character, then that character is usually presented as a question mark, a square box or another substitute character. At this time, there is no single font that supports all Unicode characters. What that means for multilingual ediscovery is that a variety of standard fonts must be supported within the processing system to ensure that text-based documents appear the way they were intended.

How Does the Litigation Team Collect Multilingual Data?

Considering today’s global business environment, whether you collect documents from custodians in the United States, Russia, France or China, there is a strong likelihood that the collected data may contain multilingual documents. Among other things, your litigation team will have to consider both the legal roadblocks set by individual countries as well as the review tools’ capabilities to handle foreign text. Some of the main international data collection considerations—location, people, tools and laws—are summarized below.

Location: The first issue a litigation team will need to consider is where the data is located. In many cases, the data is located in various countries and consists of multiple languages. Understand that for each location, you may need a separate plan of attack for gathering the data.
People: The next challenge you will face is transporting your collection team to the various countries where the data is located. Be sure to send only qualified and experienced people to work on the collection, especially when the collection occurs abroad. If possible, look to leverage a service provider with local collection experts.

Tools: Certain collection tools that are commonly used for collecting English-language data do not support collecting data in other languages. You will want to ensure that your software and hardware collection tools support any languages or character sets you might encounter in the collection.

Laws: Litigation teams must recognize that collecting data on foreign soil may raise legal roadblocks not common in the United States. Before you deploy your collection team, it is prudent to confer with local counsel to fully identify local data transfer rules, privacy regulations and jurisdictional issues specific to that location. For example, the data privacy laws in the European Union (EU) and many Asian countries attempt to protect the rights of individuals through comprehensive legislation, which may limit transport of data outside the country or require an individual to provide written agreement prior to having their data collected and examined. As far as the EU is concerned, the EU and United States have developed a “safe harbor” in which organizations can certify to the U.S. Department of Commerce that in the course of their business, they will provide adequate privacy protection when transferring personal data outside the EU.

What Should a Litigation Team Know About Filtering and Processing Multilingual Data?

Filtering and processing multilingual data is also going to be more complex if you are looking to find specific custodian names with non-Latin characters, running a keyword list that contains multilingual characters or attempting to interpret multilingual metadata fields. As with multilingual data collection, there are certain key issues your litigation team will need to consider when filtering and processing data that contain words from multiple languages and characters from multiple character sets. There are four components to filtering and processing non-English languages—character encoding, language identification, tokenization and reporting.

Character Encoding
Character encoding allows a computer system to recognize characters. While it is important for the processing system to be Unicode-compliant, there is still a need for the discovery processing system to support the most common code pages for identification of non-Unicode characters. Many ediscovery service providers tout their “Unicode compliance,” indicating that the service provider’s computer system recognizes Unicode characters. Unicode compliance alone does not mean that the application or computer system can filter or process Unicode data and allow full search ability and sort ability.

Language Identification
Another integral component of ediscovery processing is language identification. A litigation team needs to not only see the correct characters on the page, but also know what languages are included in the collection. Because an individual custodian may possess data in more than one language, it is extremely helpful if the service provider can identify both a “primary language” and a “secondary language” and pass this information along to review via a metadata field in the review tool. If the ediscovery collection includes non-Unicode data, experienced ediscovery service providers automatically identify the characters and languages of a document and convert characters as needed to a standard Unicode format.

Tokenization
Linguistic parts of speech vary substantially for different language families. Some Asian languages, for example, differ from Western European languages in the way they use spaces between words. While many languages use spaces between words to identify the beginning and end of words, some Asian languages do not. Therefore, an effective ediscovery processing system must be intelligent enough to properly parse text into “words” and sentences for filtering or searching purposes by correctly interpreting word make-up among characters. In the case of certain Asian languages that do not use spaces to identify words, a great deal of research has resulted in “lexical” processing, generally referred to as tokenization. The accompanying illustration demonstrates the different use of spaces among various languages.

The first step to effective tokenization (and effective searching) is accurate language identification. Tokenization requires language-specific software programming in order to ensure accurate search hits. Legal teams should understand how a service provider’s tools will tokenize the data set and which language structures will be used for search tokenization.

When creating keyword search lists, it is important to consider in what languages the documents may appear and what locale-specific considerations are present on the computer operating system (e.g., date/time format, currency format, etc.). Depending on the data composition, keyword lists may need to include keywords and phrases from multiple languages and additional hardware will be required (such as a keyboard for Japanese, German, etc.).

**Reporting**
One last consideration is the level to which reporting about the make-up of the data set will be available. It is important to understand which languages are contained within the documents as early as possible in a project. Look for a service provider that can generate a series of reports providing information about the languages found, document count for each language, total volume of non-English documents and language breakdown by custodian.
Know What Questions to Ask About Multilingual Processing.

A part of increasing your own familiarity with international ediscovery is knowing what questions to ask ediscovery service providers.

Good questions to ask service providers about their multilingual processing capabilities:

» What Unicode standards are your systems compliant with for character encoding?
» What happens if data is not in a Unicode format or cannot be recognized as Unicode?
» Can your processing systems search for multilingual terms? Which languages?
» Can your review tool search for multilingual terms? Which languages?
» Can the processing system and review tool identify document sets by language?
» Can you provide reports indicating the languages present in the collection?
» Do you offer human or machine translation services for multilingual data?

Generic questions about multilingual data processing that may lead to ambiguous answers:

» Can you handle foreign language data?
» Is your ediscovery service Unicode-compliant?
» What do you know about handling foreign languages in discovery?

As the common law, best practices, technology and general knowledge in the industry evolve over time, handling multilingual data in legal discovery will become less daunting. Until then, align your legal team with experienced service providers and technology consultants who can ease the burden and ensure smooth sailing through international discovery waters.

How Does a Litigation Team Review Multilingual Documents?

Even if the legal team masters the collection and processing of an international ediscovery project, challenges may still abound in the review phase. While the review team will likely follow a similar review manual to identify responsive, non-responsive and privileged documents, the legal team must decide whether and when to translate the multilingual documents into English for review. Legal teams have two options:

» Review the documents in their native languages using native-speaking attorneys—or—
» Translate the documents into English for review by English-speaking attorneys.

Often a hybrid option will be used, combining both a native-speaking review and a translated review on various subsets of the data. The issue of whether your team decides to translate first or review first will likely depend on a number of factors, many of which are outlined below.

Even a relatively small international litigation or investigation could involve more than 100,000 documents for review—thousands of which could contain languages other than English.
Reviewing in the Native Language

Documents can be reviewed before they are translated by employing a review team that is fluent in the language of the data set. This option may be expensive because it may require hiring attorneys who are fluent in multiple languages. This option may also require additional time to find qualified, experienced attorneys. For the large quantities of documents that are often involved in ediscovery projects, this expense may be overwhelming. However, a review of the documents in their native language is generally the most accurate and may be worth the cost if the documents are few, important to the case or the case involves a high risk for the client.

Translating Before Review—Machine Translation

If multilingual documents are to be reviewed in English, they must be translated first. Translation can be done by people (human translation) or computers (machine translation).

Machine translation results in a “literal” word-for-word translation; in other words, it substitutes one word for another and does not consider context or attempt to make sense of the “whole.” For example, a literal English translation of the German word “kindergarten” would be “garden of children,” but in English the phrase refers to the school year between pre-school and first grade. Machine translation does not accurately translate slang, colloquialisms or “turns of phrase” common to every language and dialect. As such, while machine translation is becoming increasingly sophisticated, it is not intended to replace human translation.

Some machine translation technologies provide for batch translation, allowing a reviewer to submit a batch of documents for translation and return to other review activities while the computer executes the translation request. Also, many machine translation technologies allow for word and phrase translation, providing an instant “on-the-fly” translation of a key multilingual word or phrase encountered during document review.

As human translation is more expensive and time-consuming, machine translation may be a good option when there are a large number of documents to review or the review needs to be done quickly. Machine translation is particularly effective when formal or formulaic language is used. For example, machine translation of legal or scientific documents more readily produces usable output than text containing casual conversation containing slang and geographic-specific vernacular.

When choosing a machine translation service, look for a service provider who supports machine translation for documents containing multiple languages in addition to supporting the primary language in the document set. Another good feature to look for is the ability to visualize multilingual data in multiple ways to suit the needs of your review: original multilingual text only, translated text only and a comparison view of translated and original texts. Often, the optimal review is achieved by utilizing a hybrid system where documents are first translated by a computer and the important documents (such as those to be offered as evidence) can be more easily identified and marked for human translation as a method of quality control.
How Does a Litigation Team Produce Multilingual Documents?

In most cases, the language of your documents does not change your duty to produce those documents. Fortunately, multilingual data does not present many new issues once you reach the production stage, although special consideration should be paid to redacted documents. The legal team must still ask the same questions regarding production format and time lines. When producing multilingual documents, four primary considerations arise: production language, production format, production ordering and sequencing and redactions.

**Language:** If you have translated the discovery documents, you have the choice of producing the documents in their native language, in the English version or both. Opposing parties can often resolve this early on in the meet-and-confer conferences.

**Format:** When you produce responsive documents, the data should be in a format that is reasonably usable by the opposing party. Production format options for multilingual documents are the same as other electronic data—native, load file, online repository and print (or some combination of these formats). However, each production format involves separate considerations:

- If producing natively, the requesting party must have software necessary to recognize the language characters contained in the data set.
- If producing a load file, you should ensure that your (and possibly the requesting party's) litigation support software can handle multilingual characters. The file may not display properly or be searchable if the litigation support software cannot display or index Unicode.
- If producing in an online repository, you should try to use the same tool used for review so that all characters are properly identified and searchable as they were for review.
- If producing printed documents, ensure that the printers have the proper print drivers installed so that all characters will be represented correctly on paper.

**Ordering and Sequencing:** When producing multilingual data, the legal team must decide whether documents should be intermixed by custodian, notwithstanding the primary language of the documents, or if the documents should be produced in a manner organized by both language and custodian. This is another discussion point that counsel will want to raise early in the meet-and-confer conferences if they are aware that multilingual data will play a role in discovery.

**Redaction Considerations:** When producing redacted documents in their native language with searchable text, it is important to verify that your service provider offers optical character recognition (OCR) support for languages included within the document production. You will want to verify that any non-redacted text is OCR’d in the redacted document’s identified native language. OCR output from redacted documents with disparate language families may contain incoherent text.

Taking stock of these issues early on will put you in the best position to avoid common international ediscovery pitfalls.
Conclusion

In our global marketplace, the volume of documents and correspondence generated between a company in the United States and its foreign subsidiaries is staggering. This international data surge has vast downstream ramifications should wrongdoing ensue and that data become relevant to a litigation or investigation. Further, various U.S. federal compliance statutes—such as Sarbanes-Oxley, the Foreign Corrupt Practices Act and others—often require American companies doing business internationally to gather and evaluate data from other countries, often in multiple languages.

As lawyers, litigation support professionals, paralegals or IT staff members at multinational law firms and corporations will attest, multilingual discovery is here to stay. Take the time now to understand the obstacles these situations present, the options you can present to your clients and the opportunities available for advocacy settings. Educate yourself and your litigation team about multilingual discovery by:

» Identifying international data locations associated with your clients’ business practices
» Researching various countries’ data protection laws and determining the most appropriate methods for collecting data in those countries
» Understanding Unicode compliance for the purpose of encoding multi-byte characters in discovery processing
» Learning about tokenization for searching multilingual data for filtering and review
» Gathering more information on human-versus-machine translation during review
» Planning for production of data in multiple languages
» Partnering with an experienced ediscovery service provider who can help your legal team navigate the waters of discovery when involved in international litigation, investigations or compliance matters

Are you prepared for international ediscovery in your next case?
Appendix: Technical Terminology

The following technical terms are important to multilingual discovery.

**Encoding:** Translation between how text is stored on a computer and the presentation of that information to the user.

**Unicode:** A character-encoding format that is quickly becoming the standard for multilingual discovery.

**Code Pages:** A character encoding format that matches documents to code pages to translate data into a readable document.

**Single-Byte Characters:** Some characters are represented with a single byte of storage space to be encoded by a computer. Character storage size is relevant for how much space a document takes when stored on the computer. Typically, the 52 characters in the English language (26 uppercase and 26 lowercase), special characters present in the other Western European languages (such as ë, é, ü), standard numbers (0–9) and common punctuation marks (‘&%^%$#@!) are stored as single-byte characters. Languages with more than 256 characters (such as Chinese) cannot be represented as single bytes due to the sheer volume of characters.

**Double-Byte or Multi-Byte Characters:** Some character sets can only be represented using more than one byte of storage space when encoded by a computer. Many languages (including the character-rich East Asian languages) can be represented using these expanded sets of available values. Unicode UTF-16 is an example of a double-byte method of storing text. For text that mainly uses the basic Latin character set, it is somewhat inefficient to store data in a double-byte format.

**Locale:** On a computer, locale refers to the operating system settings for a user's language and country. The language and country identified for the computer itself is a separate setting from language support for specific documents stored on the computer. Note that the locale is not stored in a document, therefore it is very difficult for a computer to deduce it (e.g., Canadian English is nearly indistinguishable from British English or Australian English).

**American Standard Code for Information Interchange (ASCII):** This is a character set based on the English alphabet. Most modern character sets support many more characters than did the original ASCII code page, which has a 128-character limit.

**ISO 8859:** This is a standard representation of characters/languages and is an extension of ASCII, enabling representation of characters in a broader range of languages. This standard and its variants (i.e., ISO-8859-1, also referred to as Latin-1, and the Windows 1252 code page) are the most common character encodings in use in the United States today. However, Unicode and ISO 10646 have a much wider array of supported characters.

**ISO 10646:** The Universal Character Set (UCS) is defined by the ISO 10646 International Standard as a character set on which many encodings are based. There are slight differences between ISO 10646 and Unicode, with Unicode representing a higher standard. For example, Unicode adds rules for bidirectional scripts such as Hebrew and Arabic, whereas ISO 10646 does not.
About Kroll Ontrack
Kroll Ontrack provides technology-driven consulting services and software to help legal, corporate and government entities as well as consumers manage, recover, search, analyze, produce and present data efficiently and cost-effectively. In addition to its award-winning suite of software, Kroll Ontrack provides data recovery, paper and electronic discovery, document review, computer forensics, secure information services, ESI and jury consulting, and trial presentation services. Kroll Ontrack is the technology services division of Kroll Inc., the global risk consulting company. For more information about Kroll Ontrack and its offerings please visit:

Kroll Ontrack holds a Safe Harbor Certification and can assist you in your cross-border discovery with its cutting-edge technologies, customized approach to each case and its experts, technology and offices located internationally.