Multi-language e-Discovery
Three Critical Steps for Litigating in a Global Economy
Introduction

e-Discovery has become a pressure point in many boardrooms. Companies with international operations need sound legal discovery strategies that address multiple languages.

The Problem: Produce Every Relevant Document

Data silos in different languages present a unique litigation challenge. How can companies examine all their electronically stored information for documents relevant to a case?

Step One: Processing

e-Discovery starts with processing potential evidence in many formats. Language identification and encoding software integrates foreign language documents into the processing workflow.

Step Two: Review

Once potential evidence has been identified, it must be scrutinized by both automated and human processes. Advanced linguistics tools maximize accuracy while saving time for legal staff.

Step Three: Analysis

Entity extraction gives meaning to volumes of electronic text in multiple languages.
Introduction

Managing the legal risk associated with the ever-expanding global economy is a challenge facing many companies. When litigation spans national boundaries, lawyers can be flooded with thousands of documents in languages other than English – all of it potential evidence that needs to be evaluated. In nearly every boardroom, legal exposure is in question: can the company produce every document – regardless of language – required to avoid the potential harm of civil or criminal litigation?

Companies with international operations need the ability to identify, process and review multilingual documents for use in a courtroom – discovery documents. The most prevalent hurdles to producing discovery documents include the sheer volume of electronically stored information, human limitations, unidentified languages contained in electronic documents, the character complexities and programming codes associated with different languages, and the aggregate processing costs. IT leaders must look to the next generation of e-discovery solutions to ensure that all documents demanded in litigation will be identified and analyzed to enable legal personnel to review a much smaller subset of potential evidence in an expeditious manner.

This whitepaper examines how global companies are using multi-language text processing and entity extraction tools from Basis Technology as part of their next generation of e-discovery solutions. Legal professionals and the companies they represent can comply with their discovery requirements in any language, on any scale.
The Problem: Produce Every Relevant Document

It’s a seemingly simple request: produce all documents relevant to a legal matter. But meeting discovery requirements is seldom easy, and when companies are unable to comply, the consequences can be devastating – lost cases, potentially staggering penalties and damaged reputations. The technical and legal manpower required to produce every relevant document can be enormous depending on the volume of documents under a company’s control as well as their languages and formats. Think of the number of potential repositories for company documents – email, CRM systems, accounting applications, individual computers – all in structured and unstructured formats, disparate codes and code pages. During discovery, companies must search all of these possible sources of information and identify documents that are potentially relevant to the case. This search and identify phase alone can produce millions of documents that need further legal review. And that is just for single-language search and identification – how does a global company search, identify and review all of its documents in different languages?

A typical corporate e-mail account alone is expected to generate around 4.3 gigabytes of electronic data. That number is forecast to grow to 6.7 GB per year by 2011.

- The Radicati Group

The cost to identify even preliminary, single-language documents is typically in the hundreds of thousands of dollars. Add the cost for translators or legal teams in other countries to review documents in their native languages, and the project cost and scope can be mind-boggling. What is the answer to make sure all relevant documents are identified, regardless of language?

To comply with discovery demands, companies can take advantage of automated multi-language processing and entity extraction to reduce the number of documents that need to be reviewed by legal personnel. These multi-language capabilities can be embedded within a comprehensive e-discovery strategy that allows companies to efficiently manage their domestic and international discovery obligations.

In *Victor Stanley, Inc. v. Creative Pipe, Inc.*, No. 06-2662 (D. Md. May 29, 2008), Magistrate Judge Paul Grimm said lawyers need to document and defend their e-discovery strategies. You have to defend how you reduce the data you’re producing. If you reduce 300 gigabytes of data to 25, you have to be able to explain how that happened, which means you have to show how you took foreign language ESI into account.
Making e-discovery multilingual

There are arguably many steps in the discovery process. And each step can involve hundreds or even thousands of combined hours of work. Thus, the introduction of an e-discovery solution gives companies an incredible opportunity to streamline labor intensive tasks, leading to greater efficiency and significantly lower costs.

Although there is no official e-discovery model, few would disagree that once a demand has been made for evidence stored electronically, every potential document relevant to the request must be identified. The goal in this initial identification stage is to locate potential sources of electronically stored information and determine its scope, breadth and depth. Since companies inherently store larger and larger volumes of electronic information, a simple keyword search may produce thousands or millions of initial documents. When you consider the aggregate of a company’s structured and unstructured data, the identification process alone presents a staggering workload. As more materials are identified as potential evidence, more manual review is required by legal teams, which translates to huge costs and time demands. However, inability to identify potentially relevant evidence leaves legal exposure that could result in considerably larger costs.

The answer is to deploy the right applications to assist with different steps of the e-discovery process downstream from initial identification. But when dealing with multiple languages, few solutions have addressed how multilingual services fit into and enhance an overall e-discovery workflow. These critical stages of multilingual e-discovery are processing, review and analysis, as represented in the reference model below.
Step One: Processing

Once potential evidence has been identified, preserved and collected, it is processed by an information retrieval, text mining or similar application. Recognizing and segmenting the languages used in each document is an essential step in working with multilingual electronic documents. Without this basic knowledge, the discovery application cannot accurately process structured and unstructured data.

Applications can take a fully automated approach to processing unknown text using language identification services that accurately determine both the language(s) and their encoding. These services identify a single primary language in a document or multiple languages and their boundaries, recognizing a wide selection of Asian, European and Middle Eastern languages.

**Rosette® Language Identifier**

The Rosette Language Identifier (RLI) can determine which of 55 different languages a document contains. RLI enhances evidence processing by identifying:

- A single document language
- A list of different languages in a multilingual document
- The start/end boundaries of languages in a multilingual document
- What languages are contained in the document and their percentage of the overall content

The Rosette Language Boundary Locator (RLBL) – a component of the Rosette Language Identifier – brings language identification to the sub-document level by segmenting the document in regions of different languages. Applications can use language boundary location services to route the language regions to different automated or human analysis assets, or simply to accumulate statistics about and references to the languages that appear in a collection of documents.
Step Two: Review

After processing is complete, the documents must be reviewed. Full-text search technology is typically used, and at its core is computational linguistics – the automated analysis of digital text that enables it to be rapidly stored, searched and retrieved. Robust linguistics technology is essential to legal review as it enables important linguistic services, such as tokenization, lemmatization, decompounding, part-of-speech tagging, sentence boundary detection and noun phrase extraction. The reason these services are so critical in e-discovery lies in the challenges of processing human language.

Human Language is Complex

Human language is such a deep and complex topic that there are fields dedicated solely to the study of the nature of human language, including morphology, phonetics and syntax. That’s why when technology and languages converge in e-discovery, legal professionals must ensure they are armed with the best linguistics-based solutions to minimize the risk of missing critical evidence buried in documents in different languages.

When Human Language Collides with Computer Language

Analyzing, segmenting and tagging text for review is an ongoing challenge due to the complexity of human language combined with the variety of computer codes used to process text – Unicode, Code Pages, etc. Combine code variances with complex languages like Chinese that can contain more than 47,000 characters – some even overlapping with Japanese – and you have the need for advanced linguistics solutions.

Rosette® Base Linguistics – Multilingual Text Analysis

Rosette Base Linguistics (RBL) performs critical review functions such as segmentation, decompounding and part-of-speech analysis in dozens of major languages. These functions allow a more complete review of text and fewer false-positive evidence results by focusing on languages in their purest form. RBL uses sophisticated morphological analysis – working with the specific features of a given language (punctuation, words, word forms and affixes) to segment and tag Arabic, Asian and European language text. This morphological approach applies a deep understanding of languages to return far more accurate results than a statistical treatment of text.

Text Review Essentials

- **Tokenization**
  Tokenization is the process of separating text into units of meaning, called tokens. Tokens typically correspond to words but can be difficult to locate in some languages, like Chinese or Arabic.

- **Lemmatization**
  Lemmas are the normalized form found in dictionaries. Lemmatization – reducing a word to its normalized form – allows different inflected forms to be grouped. This means that a search on the term “arbitraging” will find all documents with the term “arbitrage.”

- **Decompounding**
  Compounds are found in some languages like German, where several words can be concatenated to form a single new word. For example, “Jobbörson” (employment exchange) is formed from “Job” and “Börse.”
Step Three: Analysis and Entity Extraction

The final step in a thorough e-discovery plan is the analysis of documents identified as potential evidence. This is a critical step because it is where final documents are flagged as evidence and moved into production for delivery to a court of law or arbitrator, or to legal counsel for closer scrutiny. An automated analysis involves a discovery search in which a legal team cannot possibly know all of the relevant search terms. The technology that enables a discovery search is called entity extraction. Entity extraction automatically locates important search terms based on the same contextual cues legal professionals would use. The ability to locate these entities – or units of meaning – in any structured or unstructured text is what enables discovery search to find all the information that fits a specified legal definition even when all the search terms are not known. This ability to identify and extract entities is something the current generation of e-discovery search technologies does not do.

Humans are able to recognize entities based on a word’s context. Suppose, for example, that you wanted to find all the names of people in a document – all names, even those of people you might not have previously considered. Suppose also that the document contains the quote: “Mr. John Xyzzy spoke…” Even though you may never have seen the name Xyzzy before, you would infer that Xyzzy is a person’s name. You would base that inference on that fact that the word is capitalized and otherwise matches a pattern: “Mr. [capitalized noun] [capitalized noun] [verb]…”

Entity extraction recreates in a computer this human process of applying and recognizing context. Take the word date. When that word appears in a block of text, is it an entity with one set of contextual features – as in a point in time – or it is a completely different set of contextual features which add up to food? There are literally hundreds of different features that indicate when the word date is more about time or about food. Entity extraction technology examines these linguistic elements to determine the correct context that enables a more accurate discovery search.

Entities

Unique terms or phrases – names, places, dates, and identifiers – that give meaning to text.

How Entity Extraction Works

Words derive much of their meanings from the visual context in which they appear on a page. The same word will often have very different meanings depending on various visual cues. Similarly, the same meaning will often be expressed by very different words, also depending on visual cues. Many cues also work both visually and verbally. These cues, or contextual features, used by entity extraction technology include:

- Proximity to other words
- Written forms of the word (e.g., abbreviations, capitalization)
- Parts of speech (e.g., is the word used as a subject, predicate, object, etc.)
- Punctuation
Next Generation e-Discovery
Multi-language Processing & Entity Extraction

**Rosette® Entity Extractor**

Rosette Entity Extractor (REX) is an entity extraction technology designed to integrate with applications that classify, manage, analyze and mine textual information. Its ability to establish context in multiple languages provides a level of accuracy not available with simple keyword matching. REX is currently available for Chinese, Japanese, Korean, Arabic, Farsi, Urdu, Dutch, English, French, Italian, German and Spanish, with additional languages under development.

REX locates generic entities such as “Vice President” and “earnings estimates” as well as specific references such as “President Barack Obama” and “May 22, 2009.” This is a critical step in the process of reviewing and extracting important information from documents, preparing the information to be further structured and analyzed by other applications or legal teams.

The Impact

In 2009, legal actions against companies are more common than ever before. Boardroom decision-makers must ensure that their people, processes and technologies do not leave them exposed when producing documents for legal submission.

Discovery challenges are magnified as business operations become increasingly global, with the potential for litigation – both civil and criminal – in multiple languages. Companies that arm their legal and IT teams with superior linguistics technologies are able to perform highly accurate e-discovery at a moment’s notice to minimize legal risk to the company and its shareholders.

**The Benefit of Rosette e-Discovery Solutions**

As global discovery demands continue to grow, companies need a comprehensive strategy for producing all of the electronically stored documents that are required for legal discovery. The ability to identify, process, review and analyze relevant documents in multiple languages is critical to that effort. Basis Technology’s Rosette solutions provide multilingual capabilities for identifying the language of text, producing a normalized representation in Unicode, and locating names, places and other key concepts relevant to a legal matter. Using Rosette solutions, legal teams and IT staff can prepare effectively for current and pending litigation involving documents in virtually any language.