

Litigation Technology

A SPECIAL REPORT

BUSINESS



Don't Let Your Documents Get Lost in Translation

You must ask the right questions when choosing an e-discovery vendor to handle material not written in English.

BY JOHN FUEX

In U.S. litigation involving multinational corporations, documents in languages other than English frequently are subject to discovery demands. The technical complications of this force litigators to become more proactive in assessing the linguistic capabilities of their discovery management technology and services partners.

During the vetting process, it is not enough to ask "Can you handle foreign languages?" or single out a single aspect of the issue by asking, "Do you support data stored in the Unicode standard?"

The problem is that those who are least knowledgeable about working with multilingual documents are the most likely to overestimate their capabilities. Nonspecific inquiries like these encourage the stock industry answer "Sure, no problem!," providing scant information to differentiate among prospective legal software and service providers.

The following points should help to clarify the jargon associated with multilingual discovery and provide some guidance for interviewing vendors.

What is Unicode, and why am I hearing so much about it lately?

It has become common to find references to Unicode on the lips and in the press releases of litigation software and service providers. Often, however, the significance of the technology behind the jargon is glibly under-explained or described in nauseating detail. Unless you are a technophile or masochist, details such as the differences between Unicode variants just aren't relevant.

What you need to know about Unicode:

- Unicode is the de facto standard for translating characters and symbols of written language—both English and other languages—into numerical values for processing on computers.
- Software that doesn't support Unicode may work on Unicode

documents in English, but it most likely won't work on documents in other languages.

- Using noncompliant software on Unicode documents may cause incorrect display of non-Latin characters and sometimes data-file corruption.

That last point is why you should care about Unicode. If, for example, Japanese characters cause the review tool to omit documents from search results or improperly display them during a privilege review, an inadvertent disclosure could result. Further, a noncompliant tool can mangle Unicode characters when exporting documents for production by substituting symbols for unrecognized characters because the software didn't know how to handle them.

It is prudent to inquire about the specific Unicode capabilities of your software vendors and demand similar due diligence from service providers. Questions to ask about Unicode support:

- Will the application open and faithfully display documents in any Unicode-supported language?
- If the application creates data or document files, can it store those files using the Unicode standard?
- Can data be entered and done so by using non-Latin Unicode characters?
- Can the application handle documents with non-Latin Unicode characters in the filename?

Is Unicode the only data format I need to be concerned with?

On the fringe of the Unicode hullabaloo lurks an issue that many vendors are loathe to discuss. Despite the widespread usage of Unicode in modern productivity software such as Microsoft Office, data still lurk on network servers and backup tapes in legacy encoding formats. For example, WordPerfect still uses a proprietary encoding system for

text in languages other than English, and it is not uncommon for older documents, especially those in Asian languages, to use the legacy “code-page” technology instead of Unicode.

The variety of encoding types makes it difficult for providers to truthfully make blanket statements that they can handle all encoding formats. Although it may be unreasonable to demand an unequivocal pledge of support from vendors for any possible encoding format that might exist in your document collection, you shouldn’t let them completely off the hook. Here are some questions to guide the discussion:

- Can the solution, either software or services, handle documents in encoding formats other than Unicode (e.g. Big5, GB, or HZ)?
- Will it detect and flag documents with unsupported encodings for further analysis?
- What is your standard procedure for dealing with unsupported formats? Are there any additional charges?

Is Unicode compliance the same thing as universal language support?

No, Unicode compliance means only that the software has the ability to handle documents in languages that include characters beyond the A-Z scheme used in the Latin alphabet. The complexities of searching and reviewing a multilingual document collection are numerous and may require advanced functionality offered in very few of the available litigation tools.

Here are common linguistic complications that Unicode won’t solve and that may need specialized tools or expertise.

Compounding: Some languages, including German, Dutch, Swedish and Finnish, use compound nouns that may complicate searching. For example, without the proper search syntax, a search based on the German word *Kontaktlinse* (contact lens) would miss a document that included the word *Kontaktlinseverträglichkeitstest* (contact lens compatibility test). Specialized tools exist to facilitate searching individual components of compound nouns, but few litigation support tools have incorporated such technology.

Tokenization: To facilitate rapid searching on large document collections, search tools use a tokenization process to identify discrete words and add them to a searchable index. For most Asian languages—which use very little punctuation, don’t insert spaces between all words, and can have the meaning of characters change based on context—the process for breaking down documents into individual words can be very complex and require language-specific dictionaries. Again, few litigation tools are sophisticated enough to accommodate the idiosyncrasies of some languages.

Canonicalization: In most languages, there are multiple ways to express a single concept. Most search engines are good at handling the most common form of this in English, the synonym. Other languages, however, have more complex systems for representing concepts in multiple ways. For example, the meaning behind a Japanese ideogram can also be “spelled out” in one of several different kana character sets or transliterated phonetically into the Latin alphabet using the romaji system.

Problems also arise from languages where nouns can take on prefixes or suffixes based on the context in which they are used. For

example, in Arabic the word for “my apple” and “your apple” are represented by distinctly different canonical forms with the same fundamental meaning.

What role does automated document translation play in discovery?

Most industry experts recommend using search experts fluent in the languages present in the document collection. This is good advice and quite reasonable for certain phases of the process such as creating search-term lists for culling, reviewing documents, and final quality control. But having a translator shadowing everyone on the litigation team to translate every search isn’t always practical, especially if more than one foreign language is involved.

Machine translation can help. Although notoriously inaccurate compared to a manual process, less-expensive machine translation still can assist litigators in situations where it is impractical to have a human translator standing at the ready. Although it is not advisable to conclude

Having a translator shadowing everyone on the litigation team to translate every search isn’t always practical.



definitively that there are no relevant documents based on only a search of machine-translated versions of documents, it is quite reasonable to use automated translations to make first-pass culling decisions. For example, even a poorly translated document should provide sufficient context to discern “There’s cake in the break-room!” from “I want to report my supervisor for sexual harassment.” Of course, when significant confusion exists in the automated translation, it is best to request manual translation.

Overall, in dealing with foreign languages in discovery, it is important to take a proactive stance, be knowledgeable about the complexities, and ask the right questions of vendors early in the process to avoid costly mistakes in the discovery management process.

John Fuex is a senior discovery management consultant at IE Discovery Inc. in Austin, Texas. He can be reached at jfuex@iediscovery.com.