A Call for Processing and Search Standards in E-Discovery

Sean M. McNee, Steve Antoch
FTI Consulting, Inc.
925 Fourth Ave, Suite 1700
Seattle, WA 98104 USA
{sean.mcnee, steve.antoch}@fticonsulting.com

Eddie O`Brien
FTI Consulting, Inc.
50 Bridge Street, Level 34
Sydney, Australia 3000
eddie.obrien@ftiringtail.com

ABSTRACT

We discuss the need for standardization regarding document processing and keyword searching for e-discovery. We propose three areas to consider for standards: search query syntax, document encoding, and finally document metadata and context extraction. We would look to encourage search engine vendors to adopt these standards as an optional setup for the application of e-discovery keyword searches. We would encourage search engine users to apply these standards for e-discovery keyword searching.

Keywords

E-Discovery, search, engine, keyword, standards.

1. INTRODUCTION

E-Discovery document analysis and review continues to consume the bulk of the cost and time during litigation. As the e-discovery market matures, clients will have increased expectations about the quality and consistency of how their documents are collected, processed, and analyzed. It is also our assumption that e-discovery vendors will compete based on the quality and breadth of their review and analytic services offerings.

Seeing this as the changing landscape of e-discovery, we propose in this paper that the vendors of e-discovery software and services are encouraged to create and apply a set of shared e-discovery standards for document processing and keyword search. We hope that these standards would be organized and maintained by a standards committee such as the Sedona Conference [1] or follow the example of the EDRM XML standard.

2. AREAS FOR STANDARDS

We think there are several areas where consistency, speed, and quality could be improved by having an open and agreed to set of standards.

2.1 Search Query Syntax

Different information retrieval/search engine systems use different and often incompatible syntax to express complex searches. This can cause confusion for attorneys, for example, when they are negotiating search terms during Meet and Confer, or when they are trying to express a complex query to an e-discovery vendor.

Examples of some difficulties worth noting:

- **Wildcard operators**. Should such operators match on 0 characters or not? For example, would (Super*FunBall) hit on both the SuperFunBall and SuperHappyFunBall, or only the latter?
- **Stemming and Fuzzy Searching.** Different IR systems provide support for different algorithms for term stemming and fuzzy searching (e.g. Porter stemming or Levenshtein distance). Attempting to standardize them might be too difficult in a standard. This would be an example of a value-add that a particular vendor could offer, but only of the lawyer understand and approve it.
- **Morphology and Word-breaking.** Concepts and word breaks are hard to determine in some languages. For example, Arabic has many ways to express a single term; Chinese and Japanese have ambiguous word boundaries.

These are only a few examples of the potential problems encountered when standardizing query syntax.

Our goal here is not to suggest that any given syntax is better than another. Nor is it to "dumb down" syntax by removing extremely complex operators. Rather, we see it as a chance to set a high bar as to what lawyers can expect from search engine systems in an e-discovery context. It is quite possible that some systems simply will not have enough functionality to support a standardized syntax. In this case, the lawyers are better off knowing of this limitation before e-discovery begins!

While the syntax varies by vendor, many complex expressions have direct correlations—there should be a mapping between them. Ideally mappings would make it possible to start with a standard syntax and have each vendor map the query to their equivalent native syntax. The standard syntax should be vendor-neutral; perhaps XML or some other formal expression language should be used to define it.

2.2 Encodings and Special Characters

Textual characters are encoded in documents through the use of various character sets. The first and most well-known character set is the ASCII character set describing 127 characters (letters, numbers, and punctuation) used in English.

Lawsuits, however, are language agnostic. Unicode [2] is the preferred standard from the ISO to represent a universal character set. To state that Unicode should be used as the standard encoding for all documents in e-discovery seems obvious—so, we should do it. What is not as obvious is the need for standardized set of test documents to validate the conversion to Unicode from a variety of data formats common to e-discovery.

A position paper at ICAIL 2011 Workshop on Setting Standards for Searching Electronically Stored Information on Discovery (DESI IV Workshop)

Finally, the standardized search query syntax discussed above needs to be able to express searches for all Unicode characters, including symbols such as the Unicode symbol for skull-and-crossbones (0x2620): ☠.

2.3 Metadata and Content Extraction

A very small minority of documents in litigation are raw text documents. Most are semi-structured documents, such as emails, Microsoft Office documents, Adobe PDF documents, etc. These documents contain raw textual data, metadata, and embedded objects, including charts, images, audio/video, and potentially other semi-structured documents (e.g. a Microsoft Excel spreadsheet embedded in a Microsoft Word document).

We have an opportunity now to extend what has already been done in the EDRM XML standard to define what metadata should be considered standard extractable metadata for various file types. If we know in advance what is required, then we can ensure higher quality. For example, it will be easier to detect corrupt files. By standardizing, we also make meet-and-confer meetings smoother, as metadata no longer becomes a point of contention—both sides assume the standard is available.

2.3.1 Known Document Types

For known document types, such as Microsoft Office documents, there are several generally accepted ways of extracting content and metadata. These generally rely on proprietary technology, some of which are free (Microsoft's iFilters [3]) and some are not (Oracle's Outside In Technology [5]). Several open source alternatives also exist, such as Apache POI for Microsoft Office documents.

Relying on any one technology, whether free, paid, or open source, is dangerous. Yet, because of the complexity of these file formats, it remains a necessary requirement. By enforcing standards of what metadata and content is to be expected from this extraction technology, we can provide for a more consistent e-discovery experience.

2.3.2 The Need for Open File Formats

An important distinction for these document types is whether the file format is an open standard (email), proprietary yet fully documented (Microsoft Office [4]), or not public information. By specifying the differences between formats, a standard could enforce all data be represented in an open or documented formats. This way, open source solutions, such as Apache Tika [7], can fully participate in e-discovery without fear of reprisal. As a side effect, this could influence holders of closed proprietary formats to open them to the community at large.

One important point, however, deals with the conversion from closed to open formats. As long the standard specifies what content and metadata needs are, the conversion needs to guarantee all data comes across faithfully.

2.3.3 Information in the Cloud

For information residing in the cloud, such as documents in Google Docs, Facebook posts, Twitter updates, etc., determining what is a document can be difficult. Google Docs, for example, saves updates of documents every few seconds. Legally, how can you determine what is a user's intended save point containing a 'coherent' document?

Standardization is even more important here than for known document types—we need to define what a document even means before we can extract metadata and content. Further, all of the metadata we need might not be attached to the content but rather will need to be accessed programmatically.

3. ACKNOWLEDGEMENTS

Thanks to everyone who provided comments and insights on previous drafts of this paper.

4. CONCLUSIONS

In this paper we discussed the need for standards in e-discovery surrounding search query syntax, document encoding, and content extraction. We hope this starts a conversation among e-discovery practioners, search engine vendors, and corporations facing lawsuits with the goal of increasing search quality and consistency during E-Discovery.

5. REFERENCES

[1] Sedona Conference. "The Sedona Conference Homepage" http://www.thesedonaconference.org/, Last accessed 21 April 2011.

[2] Unicode Consortium. "The Unicode Standard", http://www.unicode.org/standard/standard.html. Last accessed 21 April 2011.

[3] Microsoft. "Microsoft Office 2010 Filter Packs". https://www.microsoft.com/downloads/en/details.aspx?FamilyID=5cd4dcd7-d3e6-4970-875e-aba93459fbee, Last accessed 21 April 2011.

[4] Microsoft. "Microsoft Office File Formats", http://msdn.microsoft.com/en-us/library/cc313118%28v=office.12%29.aspx. Last accessed: 21 April 2011.

[5] Oracle. "Oracle Outside In Technology", http://www.oracle.com/us/technologies/embedded/025613.htm. Last accessed: 21 April 2011.

[6] The Apache Foundation. "The Apache POI Project", http://poi.apache.org/. Last accessed: 21 April 2011.

[7] The Apache Foundation. "The Apache Tika Project", http://tika.apache.org/. Last accessed: 21 April 2011.