E-Discovery Revisited: A Broader Perspective for IR Researchers

Jack G. Conrad
Research & Development
Thomson Legal & Regulatory
St. Paul, Minnesota 55123 USA
Jack.G.Conrad@Thomson.com

ABSTRACT

It is a very positive development that NIST's Text REtrieval Conference (TREC) has added a track focusing on the legal (discovery) domain. Its organizers should be acknowledged for their commitment and hard work to establish preliminary tasks and arranging initial assessments. In order to ensure that the track evolves into a realistic and relevant field of study, future tracks will need to accurately reflect the nature and scope of the actual E-Discovery task, or series of tasks, at hand.

Keywords

text retrieval, legal research, electronic discovery, EDD

1. INTRODUCTION

We define EDD, Electronic Data Discovery (or E-Discovery), as any process (or series of processes) in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in a civil or criminal legal case. E-Discovery can be carried out offline on a particular computer or it can be done on a network. Court-ordered or government sanctioned inspection for the purpose of obtaining critical evidence is also a type of E-Discovery [3].

According to the Socha Report [10], the consensus among legal consumers is that 60% of today's cases warrant some form of EDD activity. This percentage will continue to grow over the course of the next several years. Regarding EDD content, according to Corporate Counsel, at least 50% of it will be in the form of e-mail, with another large chunk coming in the form of office documents (e.g., Word, spreadsheets, etc.), together with small databases (e.g., MS Access) or larger databases (e.g., Oracle), as well as less conventional forms of digitized data (e.g., software code) or other forms (e.g., voice mail or video clips) [1].

2005 represented the first billion dollar year for EDD and the market continues to expand. Interesting to note is that the top 10 E-Discovery providers cover just under half of the market, while the top 25 providers capture just over two-thirds [10]. Recognizing that this is a highly volatile marketplace, with new players arriving frequently and old players just as often disappearing or being acquired, these market share figures are unlikely to remain constant.

Given the growing reach and complexity of the field, it was fitting that 2006 marked the first year that NIST's TREC [12, 11] hosted a Legal Track [6, 5]. For the same reason, it is important that IR researchers, especially those participating in TREC's new legal track, understand just what E-Discovery entails. The goal of this paper is thus to underscore the breadth of the EDD space — which means avoiding the practice of recasting the problem as a basic retrieval task or viewing EDD as being little different from the traditional problems that West and LexisNexis have addressed. The significance of this work is that EDD does not operate on static document collections, but highly dynamic ones, and that there are a number of preliminary steps that generally must occur before the search function can even be considered. In other words, there is a vast difference between meeting the operational requirements of a production environment and baseline technology standards for research.

ICAIL'07, DESI (EDD) Workshop, June 4, 2007, Palo Alto, California USA

2. RELEVANT RESOURCES

> *There are now 300-500 vendors offering some form of EDD products or services Of those 300-500 vendors, many will be gone. Consolidation is afoot. These may be search engines, archiving tools, document management solutions, litigation support systems and more. Some offer licensed software, others sell EDD as a service.*
>
> The Socha Report & Law.com (2006)

A principal location for E-Discovery resources is the DiscoveryResources.org Web site hosted by Fios [8]. Other useful sites include Law.com [2] and the Sedona Conference [4]. A brief tutorial on the subject can be found in Barosocchini's primer [7]. An indispensable summary in the area of E-Discovery is the Socha Report [10]. This comprehensive work surveys roughly 50 EDD software and service providers and at least that many EDD software or service consumers. Some of the topics it tracks in the burgeoning field address, for instance, providers and consumers' views about EDD growth areas, about current EDD strengths, and about current EDD weaknesses. Another forecasting tool for EDD includes the "Forecast for EDD" [9].

What many of these resources explain is that E-Discovery represents a multi-stage process of materials gathering, restoring, migrating, converting, indexing, searching, analyzing and reporting. Figure 1 attempts to capture the essential features of this integrated and complex process.

3. THE EDD PROCESS

The general EDD process consists of six to eight stages, depending on the particular focus and segmentation. In some application models, the Hosting function is separate from the Search function (see Figure 1's dotted rectangle). The essential steps are described here:

1. **Identification of Content and its Scope** — breadth and depth of relevant materials identified.

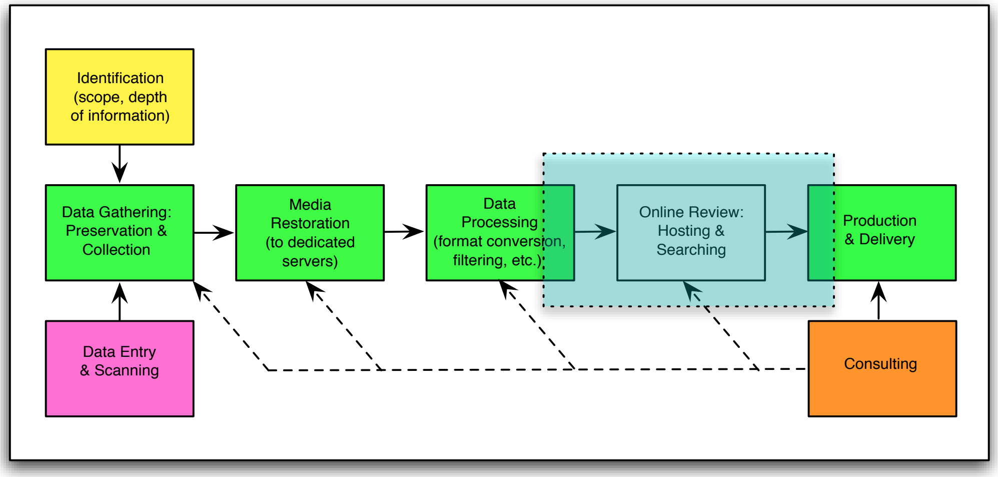
Identification (scope, depth of information)
Data Gathering: Preservation & Collection
Media Restoration (to dedicated servers)
Data Processing (format conversion, filtering, etc.)
Online Review: Hosting & Searching
Production & Delivery
Data Entry & Scanning
Consulting

Figure 1. The EDD Process — Multi-Stage Pipeline

2. **Data Gathering: Preservation & Collection** — electronically stored information is preserved from a variety of sources (e.g., tapes, PCs, networks, portable storage devices, etc.) and through a number of means.

   May include initial conversion of hard copy media (e.g., via OCR) or transcriptions of audio or video depositions or other evidence.

3. **Media Restoration** — data is transferred from original or intermediate media to uniform media on which analysis is to be performed.

4. **Data Processing** — comprises vetting task to reduce sheer volume of data.

   May include deduplication, filtering, and other means of culling or organizing via clustering or classification.

5. **Online Review: Hosting & Searching** — this is the primary stage during which data is reviewed and analyzed.

   Hosting – data may be transferred to a dedicated functional repository or inspected locally.

   Searching – can be on the basis of sources, dates, original file types, key words, concepts, etc. Data may be reviewed in a standard format (e.g., pdf, tiff, etc) or in a native format.

6. **Production & Delivery of Results** — consists of the generation & delivery of reports to varied recipients (e.g., firm associates, partnering law firms, corporate legal counsel or other service providers).

   May include delivery of packaged data for automated Litigation Support Systems. May include delivery in a variety of media forms (e.g., CD, DVD, tape, hard drive, ftp, paper).

7. **Subsequent Consulting** on the part of the Service Provider — advice to customers on procedures for conducting E-Discovery processing as well as strategies for record retention, preservation, contingency planning, etc. This stage can and often is distributed in parallel with a number of other key EDD phases.

One may assess the relative importance of these stages by using several metrics, including frequency of use by customers, the number of cases processed by each, perceived importance to customers, profitability, strategic value to a complete EDD workflow solution and others.

Another key point to emphasize is the *negotiated exchange* process that takes place between the two sides in many complex lawsuits, where what is "discoverable" is hotly contested, typically by the defendant, and often under the close supervision of the court.

4. THE TREC EXPERIENCE

Although the TREC Legal Track has only just completed its first year, the general model it has relied upon is that of the traditional text retrieval paradigm—a self-contained repository against which researchers submit Boolean or other types of queries. This appears to be a reasonable first step, since broadening its focus, even incrementally (e.g., beyond the dotted rectangle in Figure 1), can be truly daunting.

There are potential deficiencies in depicting the problem this way in the longer term, however (i.e., as a "canned" collection). After all, how can one talk about the "total recall" in an end-to-end system when one has little or no grasp of the gathering, migrating, and transformation stages that have preceded the search and analysis processes?

Furthermore, it will be increasingly beneficial to have diverse teams of researchers viewing and working on such AI & Law-related problems as cross-discipline challenges. After all, success for many applications and enabling technologies often stems from the creative generation of integrated components, hybrid solutions, and resultant synergies between those components, for example, by learning to "piggy-back" off of the filtering, indexing, or analysis of other stages.

It is also instructive to be able to go beyond these data processing stages; moreover, the track could benefit from another perspective on the EDD process, one that is less focused on the sequential nature of the problem and more focused on the technological underpinnings of the discovery process. That is, it may be a bit more intuitive to address the EDD space in terms of a technology progression or "pyramid" (see Figure 2).

The pyramid design relies upon the following structures:

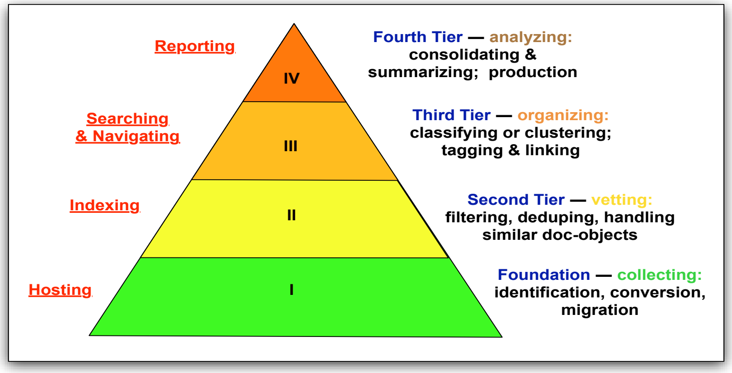

Figure 2. The EDD Pyramid — A Technology Perspective

1. **Foundation Tier (Hosting)** — *Collecting:* identification, conversion, migration.
2. **Second Tier (Indexing)** — *Vetting:* filtering, deduplication, managing similar objects.
3. **Third Tier (Searching & Navigating)** — *Organizing:* classifying & clustering; tagging & linking related documents.
4. **Fourth Tier (Reporting)** — *Analyzing:* production: consolidating & summarizing findings.

The distinct advantage of viewing EDD in terms of such a pyramid includes (a) not being limited by the constraints of a sequential pipeline, and (b) being able to build upon the foundations established in the preceding tasks (e.g., indexing following on the heels of hosting). This alternative model may be more suitable for researchers attempting to tackle difficult precision and recall problems in contrast to engineers expected to satisfy operational constraints while moving the pipeline closer to production. As the descriptions above indicate, the model is focused principally on the building blocks of text processing and retrieval tasks which are keys to higher performance systems and the delivery of their results to front end users.

5. CONCLUSIONS AND RECOMMENDATIONS

As NIST focuses on contributing to the state of the art in Electronic Discovery, it will be imperative to have the problem space accurately reflect the practical considerations that the legal field is currently facing. This includes the proper recognition of the stages involved in EDD, or, alternatively, the tiers of related technology areas. In order to permit researchers to effectively address the challenges of the evolving field, it will be helpful to give them exposure to the numerous practical considerations that those working in the commercial sector currently benefit from. This might also entail engaging commercial enterprises to assist in expanding the current thrust of the E-Discovery track from one that focuses largely on restricted search to one that examines the interaction of parallel research areas, like those depicted in Figure 2. The more IR researchers know about the broader context of E-Discovery, including the stages that come both before and after core retrieval activities, the greater the prospects for broader solutions, creative optimizations and synergies yet untapped.

6. ACKNOWLEDGMENTS

We thank Peter Jackson and Khalid Al-Kofahi for the time and resources to pursue this study.

7. REFERENCES

[1] Corporate Counsel. The American Bar Association (ABA), Section of Litigation, Committee on Corporate Counsel, 2006. www.abanet.org/litigation/committees/corporate/.

[2] Law.com. Web-based legal news and information network: www.Law.com, 2007.

[3] Search Security. IT site to keep corporate data and assets secure: www.SearchSecurity.com, 2005.

[4] The Sedona Conference. Facilitates discussion among legal experts on topics like complex litigation: www.theSedonaConference.org, 2007.

[5] Jason Baron and Paul Thompson. The search problem posed by large heterogeneous data sets in litigation: Possible future approaches to research. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAIL07)*, Palo Alto, CA, June 2007. ACM Press.

[6] Jason R. Baron, David D. Lewis, and Doug W. Oard. TREC 2006 legal track overview. In *The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*, Gaithersburg, MD, Nov. 2006. National Institute of Standards and Technology, NIST.

[7] Albert Barosocchini. Electronic discovery primer, 2005.

[8] Fios. Discovery Resources Web site. Resources and news about E-discovery: www.discoveryresources.org, 2007.

[9] Alan Radding. The forecast for EDD, Nov. 2006.

[10] George Socha and Tom Gelbmann. The 2006 Socha-Gelbmann Electronic Discovery Survey Report. Report, Socha Consulting, Saint Paul, MN, 2007.

[11] Ellen M. Voorhees. Overview of TREC 2006. In *The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*, Gaithersburg, MD, November 2006. National Institute of Standards and Technology, NIST.

[12] Ellen M. Voorhees and Lori P. Buckland, editors. *The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*, Gaithersburg, MD, Nov. 2006. NIST.