## ICAIL 2007 Workshop, June 4, 2007, Palo Alto, California

*Title:*

# Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings ("DESI Workshop")

*Background*

Lawyers and their large institutional clients increasingly face the enormous problem of how to efficiently and efficaciously conduct searches for relevant documents in heterogeneous haystacks of electronic data. The heterogeneous complexity of datasets subject to discovery is rapidly approaching the threshold of where hundreds of millions of documents are being made subject to more or less "routine" searches in a variety of litigation and investigatory contexts. In recognition of this phenomenon, as of December 1, 2006, the U.S. Federal Rules of Civil Procedure officially recognized "electronically stored information" (ESI) as a term of art embracing all forms of electronic documents made subject to the civil discovery process. Under Rule 26(f), opposing parties to federal court litigation now have an early "meet and confer" duty to discuss a range of issues, including the continued storage, preservation, and access to ESI in their respective physical and legal custodies. As one part of these discussions, courts can be expected to envision and require that parties make good faith efforts to collaborate on coming up with a set of search protocols and parameters used to govern the future course of discovery in the case. However, most present-day litigators find themselves ill-equipped to evaluate state-of the-art legal tech sector claims as to what search retrieval methods, tools, and techniques should be utilized. This is especially true given the state of the case law, where the leading cases do no more than discuss the need for search protocols by reference to "keywords."[1]

In particular, two initiatives have emerged over the past few years that begin to address these challenges. The first has been the work of The Sedona Conference ® working group on electronic document retention and production, composed principally of legal professionals with experience in civil discovery involving ESI, including both lawyers and so-called "e-discovery" firms.[2] The second is a new "legal track" sponsored by the National Institute of Standards and Technology (NIST) as part of its annual Text Retrieval Conference (TREC), consisting of a multi-year collaborative information retrieval research project focused on information retrieval for e-discovery applications in which both academcs and corporations participate. What is still missing to date is broader engagement with the research community on development of new technologies to support the e-discovery process. E-discovery applications often involve massive datasets that couple formal documents both in character-coded and scanned formats, informal communications from a variety of systems, metadata in both standardized and nonstandard forms, and databases. Information access in these applications involves more than simply search—helping the searcher to make sense of what they find can be equally important. Protection for confidentiality can be particularly challenging in e-discovery applications, both because a complex set of interlocking rights and privileges must be accommodated and because sensitive information is often intermixed in ways that are difficult to segregate.

Our goal in this workshop is to begin the process of crafting a broad community-wide research agenda to address these challenges. We believe that this is a propitious time for such a workshop, and that ICAIL is the natural place to begin the process. We see this as the start of a process that will ultimately grow to include specialized research in multiple disciplines, perhaps as a result of additional discipline-specific workshops.

*Organization of the Workshop.*

The full-day workshop will be organized in five parts:

Part I will begin the workshop with a keynote presentation from a leader of the Sedona Conference ® working group on electronic document retention and production. The goal of this talk will be to lay out the challenges as seen from the application community, and to begin to name the technical issues that are involved.

Part II will consist of contributed research presentations. We plan to issue a broad call for participation to every community in which the organizers are active (The Sedona Conference ®, e-discovery firms, information retrieval, human language technology, natural language processing, text classification, archives, library science, information studies, and human-computer interaction) in which we will solicit both research contributions and position papers. If the number of contributions meriting presentation exceed what can be accommodated in this section, we will add a poster session during an extended break between Parts IV and V.

Part III will be a set of moderated "breakout sessions" over lunch in which participants will brainstorm ideas for a research agenda. We plan to form groups for this session based on brief questionnaires that are collected at the beginning of Part II, and to limit the size of each group to eight people at a single table.

Part IV will be invited presentations describing available test collections. We presently know of two such collections. The full Enron collection includes about 500,000 emails (many with attachments and some with discourse and topic annotations), an extensive set of databases that focus mostly on energy trading activity, a very large set of recorded telephone calls (some of which have been transcribed, and some of which have discourse annotation), and a set of 15 use cases. A number of research projects that have used this collection will be briefly reviewed to illustrate its potential and its limitations. The Complex Document Information Processing (CDIP) Test Collection from the Illinois Institute of Technology (IIT) includes 6.9 million documents released by tobacco companies in connection with a "Master Settlement Agreement" with several state attorneys general. 32,738 topical relevance judgments are available for 40 use cases.[3] The CDIP collection was used in the 2006 TREC legal track, and results from those experiments will be briefly reviewed to illustrate the collection's potential and its limitations. We also plan to briefly describe a third collection of records from a failed law firm that is expected to become available for limited research use within a protected environment in this time frame because it provides a useful example of the confidentiality issues that arise when working with these types of materials. We will also solicit contributions describing potentially useful test collections of which we are not presently aware.

Part V will consist of a panel discussion, with one panelist for each lunch table topic (for multitable topics, the moderators will meet over the break between sessions IV and V to consolidate their ideas). The background of these panelists will be balanced to ensure inclusion of at least one member from three key constituencies: (1) an e-discovery firm, (2) a practicing lawyer, (3) academic researchers from at least two disciplines. We may also invite a program officer from the National Science Foundation to participate. The panel discussion will include brief presentations by each panelist, followed by a facilitated open forum addressing questions such as:

- Who, beyond those already in the room, do we need to engage with to address the challenges that we have identified?

- Are the test collections and evaluation measures that we are using adequate to explore the range of research questions we need to consider? If not, what new developments are needed?

- What do we know from earlier research on technical support for search and sensemaking in other legal applications (e.g., legislation and case law found in structured databases) that would provide e-discovery researchers with useful starting points?

- What research questions should be explored in the TREC legal track that have not been yet addressed there?

- How can we help to inform the professional practice of legal professionals? For example, how can we help lawyers achieve optimum results when negotiating search queries, in satisfaction of the new "meet and confer" obligations relating to ESI?

Of course, the actual questions that we address will emerge from our discussions over the course of the day. We will create both a public Web page and an open mailing list to disseminate the results of the workshop, and if there is interest among the participants we may also consider a more formal publication proposing a research agenda.

We plan to solicit contributions and participation through The Sedona Conference ®, and to the legal and research communities identified above through appropriate mailing lists. An optimal size for the workshop would be 25-30 participants, but we expect that the diverse range of research issues raised by the challenges we will address, the large number of research communities involved, the amount of commercial interest in this topic, and the location of the conference will result in many more interested participants than that. We will therefore give first preference to those who are presenting and then accept additional participants up to whatever capacity the room can accommodate.

*Organizing Committee*

Jason R. Baron
Director of Litigation
National Archives and Records Administration
8601 Adelphi Road, Suite 3110
College Park, MD 20740
jason.baron@nara.gov
tel. 301-837-1499

Jason Baron has served for the past six years as Director of Litigation for the National Archives and Records Administration. Between 1988 and 1999, Mr. Baron held successive positions as trial attorney and senior counsel in the Civil Division of the Justice Department in Washington, D.C., where he represented the interests of the U.S. government in a variety of complex lawsuits including involving access to governmental information. Mr. Baron appeared as lead counsel in two cases filed against the White House and the Archivist of the United States involving electronic records management (*Armstrong v. Executive Office of the President* and *Public Citizen v. Carlin*). Among his publications, he is the author of two recent law review articles entitled "Toward A Federal Benchmarking Standard For Evaluating Information Retrieval Products Used In E-Discovery," published in 6 *Sedona Conference Journal* 237 (2005), and, with co-author George L. Paul, "Information Inflation: Can The Legal System Cope?," published in the online *Richmond Journal of Law and Technology.* He currently represents NARA on The Sedona Conference ® Working Group on Electronic Records Retention and Production, where he co-chairs a Special Project Team on Search and Retrieval Sciences. For the past two years he has served as co-coordinator of the National Institute of Standards and Technology TREC (Text Retrieval Conference) legal track. He has been a Visiting Scholar at the University of British Columbia, an Adjunct Professor at the University at Albany, and is currently an Adjunct Professor at the University of Maryland's graduate College of Information Studies. He also presently serves on the Georgetown University Law Center Advanced E-Discovery Institute advisory board. Mr. Baron received degrees from Wesleyan University and the Boston University School of Law.

Douglas W. Oard
Associate Professor, College of Information Studies
and Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
oard@umd.edu
tel: 301-405-7590

Douglas Oard is Associate Dean for Research in the College of Information Studies at the University of Maryland. An Associate Professor in the College, he holds a joint appointment in the Institute for Advanced Computer Studies (UMIACS) and affiliate appointments in the Computer Science Department and the Applied Mathematics and Scientific Computation Program. Dr. Oard earned his Ph.D. in Electrical Engineering from the University of Maryland, and his

research interests center around the use of emerging technologies to support information seeking by end users. Recent work has focused on interactive techniques for cross-language information retrieval, searching conversational media, and leveraging observable behavior to improve user modeling. Additional information is available at http://www.glue.umd.edu/~oard.

David D. Lewis
David D. Lewis Consulting
858 W. Armitage Ave #296
Chicago, IL 60614
davelewis@DavidDLewis.com
tel: 773-975-0304

David D. Lewis is an independent consultant based in Chicago. He works in the areas of information retrieval, machine learning, natural language processing, and the evaluation of information systems. Dr. Lewis previously held research positions at AT&T Labs, Bell Labs, and the University of Chicago. He earned his Ph.D. in Computer Science at the University of Massachusetts at Amherst. Dr. Lewis has published more than 40 papers and 5 patents, and is a frequent invited speaker and tutorial presenter.

Paul Thompson
Research Associate Professor
Department of Computer Science
Dartmouth College
6211 Sudikoff Laboratory
Hanover, NH 03755-3510
Paul.Thompson@dartmouth.edu
tel 603-646-8747

Paul Thompson is a Research Associate Professor in the Department of Computer Science at Dartmouth College. He also held an appointment as a research engineer at the Institute for Security Technology Studies (ISTS) and the Thayer School of Engineering, and participates in research conducted by the Institute for Information Infrastructure Protection (I3P). In addition he is an instructor in a graduate information assurance program at Norwich University. Dr. Thompson earned his Ph.D. in Library and Information Studies at the University of California, Berkeley. He currently holds affiliate appointments with the ISTS and its Cyber Security and Trust Research Center (CSTRC). His research interests focus on mixed-initiative interaction in information retrieval, question answering, information extraction, text mining, information infrastructure protection, and deception detection. Additional information is available at http://www.ists.dartmouth.edu/cstrc/projects/semantic-hacking.php. Prior to his current appointments at Dartmouth College, he was a principal computer scientist in the AI Development Group of PRC, Inc. (now part of Northrop Crumman) from 1988-1993, and a research scientist at West Publishing, later West Group, from 1993-2001.

*References*

Baron, Jason R., "Toward A Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery, " 6 *Sedona Conference Journal* 237-246 (2005) (available on Westlaw, Lexis)

Collaborative Expedition Workshop #45, *Advancing Information Sharing, Access, Discovery and Assimilation of Diverse Digital Collections Governed by Heterogeneous Sensitivies*, held Nov. 8, 2005, http://colab.cim3.net/cgibin/wiki.pl?AdvancingInformationSharing_DiverseDigitalCollections_HeterogeneousSensitivities_11_08_05

NIST TREC 2006 Legal Track web page, http://trec-legal.umiacs.umd.edu (TREC 2006 conference proceedings forthcoming).

Paul, George L. and Jason R. Baron, "Information Inflation: Can The Legal System Cope?," 13 *Richmond Journal of Law and Technology* (2006), http://law.richmond.edu/jolt/v13i2/article11.pdf.

The Sedona Conference, *The Sedona Principles: Best Practices Recommendations & Principles for Addressing Electronic Document Production* (2005 version), see http://www.thesedonaconference.org/content/miscFiles/publications_html

ENDNOTES

[1] *See, e.g., Zubulake v. UBS Warburg, LLC,* 229 F.R.D. 422 (S.D.N.Y. 2004); *Treppel v. Biovail Corp.,* 233 F.R.D. 363 (S.D.N.Y. 2006); *Balboa Threadworks v. Stucky,* 2006 WL 763668 (D. Kan. 2006)

[2] A legal commentary on the subject of search and retrieval issues is expected to be published by The Sedona Conference ® in 2007.

[3] *See* NIST TREC 2006 Legal Track web page, http://trec-legal.umiacs.umd.edu (TREC 2006 conference proceedings forthcoming).