ARTICLE:

E-DISCLOSURE VIEWED AS 'SENSEMAKING' WITH COMPUTERS: THE CHALLENGE OF 'FRAMES'

By **Simon Attfield** and **Ann Blandford**[1]

In addressing the question of the design of technologies for the purposes of e-disclosure (this includes the term e-discovery, as used in some jurisdictions) it is essential to recognize that people and technology interact as a complex whole. Technology can promote disclosure and support the ability of people to make sense of data, but the extent to which it is able to do this depends upon the extent to which it naturally extends the way that legal practitioners think and work. This paper describes research undertaken at University College London, which uses this as a starting point for empirical studies with the intention of influencing the design of supporting technologies. A field study comprised interviews with lawyers who worked on a large regulatory investigation. Using data from this study, the document review and analysis is described in terms of a sequence of transitions between different kinds of representation. The paper focuses on one particular transition: the development of a chronology from the documents. The authors consider the idea that investigators make sense of evidence by the application of conceptual 'frames',[2] but whilst the investigator 'sees' the situation in terms of these frames, the system 'sees' the situation in terms of documents, textual tokens and metadata. The authors conclude that the design of software can be improved through the development of technologies that aggregate content around the 'frame' perceived by the investigators. Research is suggested to explore this further.

Introduction

Electronic Data Disclosure (EDD, or e-disclosure) has been defined as a process (or series of processes) in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in civil or criminal proceedings, or as part of an inspection ordered by a court or sanctioned by a government.[3]

The rapid increase in the volume of electronically stored information within modern enterprises has led to a situation in which preparing for and executing e-disclosure represents a considerable challenge for organizations and lawyers, and it is one that is set to increase. It is also possible, in the opinion of a director of a vendor selling disclosure products, that companies may have difficulty if they cannot uncover all electronically stored information (ESI) relevant to a legal or regulatory matter within a specified time, as required by a regulatory body.[4]

Advances in digital technologies have brought about this challenge, but technology also offers part of the means to address the problem. The e-disclosure technology industry is expanding. For instance, software revenues in 2006 in the United States of America were estimated at around US$150 million, with further vigorous growth predicted.[5] Technologies attracting particular interest in this arena include media

[1] *The authors thank Freshfields Bruckhaus Deringer for their kind help with the fields study reported here. The work was funded under EPSRC grant EP/D056268.*

[2] *Gary Klein, Jennifer K. Phillips, Erica L. Rall, and Deborah A. Peluso, 'A Data-frame Theory of Sensemaking' , in Expertise Out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making, edited by Robert R Hoffman (New York: Lawrence Erlbaum Associates, 2007), pp 113-155.*

[3] *Jack G. Conrad, ' E-Discovery Revisited: A Broader Perspective for IR Researchers' , DESI Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings, June, 2007, http://www.umiacs.umd.edu/~oard/desi-ws/papers/conrad.pdf.*

[4] *Deborah Baron, ' UK Firms Report Jump in Spend on E-discovery Systems' , Computer Weekly, 17 March 2008, http://www.computerweekly.com/Articles/2008/03/17/229880/uk-firms-report-jump-in-spend-on-e-discovery-systems.htm.*

[5] *George J. Socha Jr., Esq., Tom Gelbmann, ' The 2006 Socha-Gelbmann 4th Annual Electronic Discovery Survey Report' http://www.sochaconsulting.com/ 2006survey.htm.*

restoration tools, dedicated document management systems, advanced information retrieval systems (such as concept search and information extraction), information visualization and case analysis tools.

In addressing the question of how to design technology for e-disclosure, however, it is essential to recognize that e-disclosure takes place as a process in which people and technology interact. In this context, the role of technology is to provide tools and resources that can be used by legal professionals, often working in teams, in constructing strategies and processes that enable them to undertake their work more effectively. Understanding how technologies can offer additional support, depends on how the technologies affect and reshape such systems for the better.

In considering the development of technology, a significant research objective is to view the e-disclosure process as a complex system of work. More specifically, e-disclosure can be thought of as a form of sensemaking activity. ‘Sensemaking’, is a topic of research that has developed increased significance in areas such as Information Science and Human Computer Interaction in recent years. Sensemaking is what people do when they face new problems or unfamiliar situations, and their current knowledge is insufficient.[6] Importantly, sensemaking typically involves more than finding information. It can also includes gathering information, re-representing it in a way that aids analysis and insight, and performing some action based on the insight.[7] Such a perspective becomes particularly pertinent where people are required to engage in intense cognitive activities such as information assimilation, theorizing and reasoning, as occurs during the review and analysis of large collections of documents. Technology can promote sensemaking in e-disclosure and support people when working on the task, but it is necessary to ensure that technology integrates with and naturally extends the way that legal practitioners think and work.

It is suggested that the design of systems to support this kind of work needs to be predicated upon an understanding of the cognitive and social aspects of e-disclosure in practice. This requires a detailed understanding of the task as it unfolds, including the associated processes of making sense of the materials, teamwork, how people currently use different tools and resources to meet their aims, and the barriers and difficulties that arise in doing so. In essence, the need is to examine how the work is undertaken in order to speculate how it might be improved upon.[8]

With this in mind, the authors are in the process of conducting research in the field and the laboratory, with the aim of more fully understanding how the review and analysis of evidence is conducted in e-disclosure exercises in order to support the design of supporting technologies. This article explains the results of the work relating to an interview field study performed with lawyers who worked on a large regulatory investigation.

In analyzing the data from this study, two complementary perspectives emerged. The first, which is reported elsewhere,[9] focuses on how the investigation work was structured. This concerns how the investigators divided it into multiple lines of enquiry that emerged in response to continuing discoveries and how these were distributed across the team. Significant issues arose in relation to how findings from multiple threads of the investigation were ultimately integrated to form an overall perspective of the case.

The second perspective is that of process. Work that involves the manipulation of information, or knowledge, often occurs in stages. Further, these stages take the form of a series of transformations between different kinds of intermediate representation or work product.[10] These representations can be the objectives of the task (such as questions), search results, or findings and interpretations (such as notes and narratives). As a representation is created or changed, so it provides the raw material for further work, creating new representations and so forth. In this way, the gradual increase in information helps to make sense of the data.

The focus is on this second perspective in this paper. An overview of the process of document review and analysis in terms of a sequence of transitions between different kinds of representational resource is first described, before focusing on one transition in detail. How this transition was achieved is described, and this is used as a catalyst for a discussion which reflects on alternative technologies that might offer additional help.

[6] *Brenda Dervin, ‘ From the Mind’s Eye of the User: The Sense-Making Qualitative-Quantitative Methodology’ , in Qualitative Research in Information Management, ed. by Jack D. Glazier and Ronald R. Powell (Englewood, CO: Libraries Unlimited Inc., 1992), pp 61-84.*

[7] *Peter Pirolli and Stuart Card, ‘ The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis’ , Proceedings of the International Conference on Intelligence Analysis, 2005 https://analysis.mitre.org/proceedings/Final_Papers_Files/206_Camera_Ready_Paper.pdf.*

[8] *Jens Rasmussen, Annelise Mark Pejtersen and L. P. Goodstein, Cognitive Systems Engineering, (New York: John Wiley & Sons, 1994).*

[9] *Simon Attfield, Ann Blandford and Stephen De Gabrielle, ‘ Investigations Within Investigations a Recursive Framework for Scalable Sensemaking Support’ Sensemaking Workshop, ACM SIGCHI Conference, 2008 http://dmrussell.googlepages.com/sensemakingworkshoppapers.*

[10] *Simon Attfield, Sarah Fegan and Ann Blandford, ‘ Idea Generation and Material Consolidation: Tool Use and Intermediate Artefacts in Journalistic Writing’ , Cognition, Technology and Work (Online First, 28 February 2008) http://www.springerlink.com/content/ll5202784t974504/.*

Investigation process

Based on the interviews with lawyers, a description of the document review process was developed, as shown in figure 1, in the form of a 'process-resource' model. In this figure, boxes represent resources and arrows represent transitions between them. For example, given a set of investigation issues and a collection of documents recovered from various locations within the firm under investigation (called a 'document universe' by the lawyers), keyword searches (t2) resulted in sets of search results. Given a set of search results, the initial manual review (t3) produced sets of documents coded as relevant to one or more 'issues' that formed the focus for the investigation.

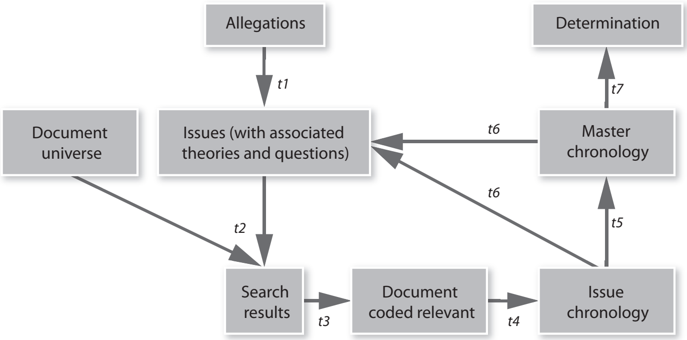

FIGURE 1

As the investigation progressed, the resources illustrated in figure 1 were continually revised. Further, given the transformations performed between resources, a change in any one had an effect on the next. For example, revised theories and questions led to revised search results, which led to revised documents coded as relevant. The transformations occurred by processing information, such as when the investigators reviewed a resource and recorded their conclusions, or by using automated processing, such as information retrieval.

Each transformation has the effect of using one or more resources in order to define another, with each representing an intermediate step that ultimately links allegation with a conclusion.

In overview, the transitions were:

t1: Given the allegations, the investigators defined and recorded a set of issues that they wanted to investigate and associated questions they wanted to ask.

t2: Given these questions, queries were submitted to the document universe to return documents relevant to each of the issues.

t3: Returned documents were individually read and coded for relevance to the issues (within a document management system).

t4: Relevant documents were then used to place entries within a chronology that was specific to each issue.

t5: Selected entries within separate chronologies were put into a single master chronology designed to record the most significant aspects of the developing narrative.

t6: By reflecting upon the narratives within the chronologies as they evolved, the investigators were able to identify apparent gaps, inconsistencies and time-periods of potential interest. This helped them to develop theories, which in turn guided the refinement of the investigation and associated questions.

t7: Given the knowledge acquired, the investigators formed a view concerning the allegations.

The structure of the investigation process evolved over time. The description provided, reflects the process in its mature form. The discussion is restricted to a description of the investigation as it applied to electronic documents, omitting reference to witness interviews which were nevertheless an important, if non-technological, source of information. A number of things occur in this process, but broadly it can be considered to be a process of information reduction achieved by different kinds of filtering and abstraction, directed by reflective interpretation on the part of the investigators.

Two things are important to note. First, the investigators constructed each step for a reason - this being in general terms to help them move in a direction in which they wanted to go. Hence it is possible to learn about their needs from what they did. The second point is that although they had discretion to design the process as they saw fit, they did so within the constraints of the tools available to them at the time, and whatever costs there were associated with their appropriation and use. Hence it is possible to use the process to consider other tools which may have supported their needs better.

Focusing on transition t4

In considering where new technologies might offer help in an e-disclosure exercise, the focus of attention might usefully be directed in detail to any part of the process described, to consider how things might be changed (or even change the process as a whole). The interviewees consistently cited manual document review (stages t3 and t4 in figure 1) as imposing the major cost in terms of time and effort. Over the course of the investigation, 130,000 documents were reviewed. This represents a significant reduction on the document universe, but is nevertheless a significant number of documents. Transition t4 will be considered in more detail.

T4 involved the creation of semi-standardized event records based on the review of documents which themselves had been coded as relevant to an issue. The investigators constructed chronologies as tables using Microsoft Excel according to a preformed schema. An example of a record that has been rendered anonymous and which reflects this schema is shown in figure 2.

Date	Time	Event / Document	People Involved / Author/Recipient	Evidence/ File Reference
8th November	7.45	{company A} meeting in {country A} (time is {person B} flight departure from {location A} to {location B}) with return to {location A} for 12.55 on 9th November {person I} to pick up {person B} at airport	{person I}, {person B} and {person H} in {location C}	E-mail between {person I}, {person I}, {person H} and {person F}/Doc ID 169246

FIGURE 2

The reason for creating chronologies was to enable the investigators to have a clear representation of the events they considered to be of significance to their investigation. This then provided a resource for considering what they knew, for developing theories of the case, and to help establish what it was they wanted to find out (transitions t6).

The resource for creating event records (transition t4) was a list of documents (predominantly e-mails) displayed in overview and listed chronologically within a document management system. The task of the reviewer was to review each document in turn and, where appropriate, create a record of any event of potential significance to the investigation. For example, this might be a meeting proposed by e-mail between two people.

An appreciation of which events were significant to the investigation (and hence what to record) evolved over time as the investigators’ understanding developed and they reviewed the theory of the case. In this paper, the focus is on what happens when an investigator first discovers information about a potentially significant event. The information contained in the message acts as a cue to the investigator about something that should be recorded. However, they are also aware that the information they have found is not the complete picture. In respect of a meeting, for instance, the investigators described a number of things they might like to know, such as where and when it took place, who attended, what was discussed and the conclusions that were reached. Some or all of this information might be missing from the initial item, and may be found distributed across a number of other messages. In addition, the initial lead may have been misleading: there may have been a change in plans or the meeting may not have actually taken place.

Following Klein’s model of the process of sensemaking,[11] the investigator’s concept of an event is described as an instance of a ‘frame’. Frames are structures that we impose on the world in the process of understanding it. They are triggered by cues and act as plausible interpretations of those cues. A significant property of a frame, which is important in this context, is that they extend beyond the data from which the cue was first formed. The ability to interpret things in this way is a fundamental human capacity. But as a consequence, they can be wrong, perhaps as a result of a misleading cue.

Returning to the investigation, following the initial discovery or cue, a question arises about how to proceed. The initial document provides an important lead, prior to which the investigator knew nothing of the event. The investigator may have a theory that a significant meeting took place. This theory gives rise to a need for further information, specifically in order elaborate and validate the interpretation.

The investigators interviewed described two strategies. Given potential difficulties in locating other

[11] *Gary Klein, Jennifer K. Phillips, Erica L. Rall, and Deborah A. Peluso, ‘A Data-frame Theory of Sensemaking’, in Expertise Out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making, edited by Robert R Hoffman (New York: Lawrence Erlbaum Associates, 2007), pp 113-155.*

documents about a given event, one strategy was simply to record the event as a conjecture and move on. Investigators would raise an event record in a chronology (marked as a conjecture) and continue reviewing documents as before in the hope that they, or someone else, would come across further relevant information later. The second strategy was to construct further keyword and date queries designed to interrogate the collection using different criteria to establish whether there were any more relevant documents relating to the matter.

Whilst the second strategy offers continuity to the investigator in terms of focus by supporting a single chain of thought, it is also a strategy which may be difficult, in practice, to carry out. The investigator sees the task they are investigating in terms of events, whilst the technology they are using structures the data in terms of documents, textual tokens and metadata. Consequently, the investigator must translate their question (of all documents relevant to a particular event) into something understood by the system - referred to more generally as a 'compromised need'.[12] This can require some cognitive effort, and mean the investigator does not obtain precisely what they want, especially if the data they are looking for is actually present amongst multiple documents in the system.

Reflections on design

This example suggests a general principle that can be applied to such problems. That is - where a user is making sense of information through the application of a particular type of frame (or frames), it is useful for the system to link any documents with information that might be relevant to that particular frame, and display these to the user as some form of linked set. Of course, there may be a number of types of frame that are important to an investigator. Other frames identified from the interviews included business activities (such as contracts), particular time periods surrounding major events within those activities, and protagonists or potential protagonists under investigation. In many of these cases, information discovered within the collection of documents gave cues to the investigators to their existence, and acted to help with further investigation of the facts (the investigators started with almost no knowledge). In all cases, further information was distributed across the document collection.

Frames that were of significance to the investigators were reflected in the way they structured the representation of knowledge they generated, that is, chronologies were set out in terms of individual events and individual chronologies were dedicated to information about specific business activities and people. Hence, in attempting to understand how the investigator wants to understand the information and the frames that they seek to apply, it may only be necessary to consider the nature of the representations they seek to create.

Construing the investigation from the perspective of the investigator, it is a question of how an investigator establishes a relevant fact, and how they continue to search for relevant data that supports the theory, or if there is no further information relating to the fact, whether the lack of any further evidence enables the investigator to eliminate the theory from their investigation. It follows from this, that is it possible to determine the extent to which information retrieval technologies support this thought process.

This question is addressed to some extent in relation to the fact that once a theory arises with its associated questions, to pursue these questions the investigator must translate their question into a query that the system can perform. Typically, this will involve translating their question into relatively basic features, such as keywords, dates, custodians and such like.

It may be useful to determine what other kinds of technologies might be developed which may be more helpful. The question here is one of additional system intelligence that might achieve the potential for aggregating documents in terms which more closely match the concepts of the investigator. In this way, transformations performed earlier in the process (such as a search) could organize the data in a way that is better adapted for subsequent work.

A number of possibilities exist. First, systems that offer representations of e-mail documents in terms of subject threads may offer some advantage. Analysis of the Enron collection, however, has suggested that the average length of an e-mail thread (in organizations at least) is typically quite short.[13] Also, systems that are capable of semantically clustering documents (such as Attenex Patterns Document Mapper[14]) may be of value, depending on the extent to which document clusters relate to investigators' conceptual frames.

Another alternative is to use systems that perform information extraction (IE). IE systems process free text and use techniques in computational linguistics in order

[12] *Robert S. Taylor, 'Question-negotiation and Information Seeking in Libraries', College and Research Libraries, 29 (1968), pp 178-194.*

[13] *Bryan Klimt and Yiming Yang, 'Introducing the Enron Corpus', First Conference on Email and Anti-Spam (CEAS), 2004, http://www.ceas.cc/papers-2004/168.pdf.*

[14] *See http://www.attenex.com/products_services/attenex_patterns_suite.aspx#Attenex_Patterns_Document_Mapper.*

to identify pre-defined elements of meaning.[15] Jigsaw, for example, is an investigators tool specifically designed to provide a graphical representation of the results of information extraction over a free text collection.[16] Elsewhere, capabilities for identifying temporal and event references in text have been demonstrated at 83 per cent accuracy against hand-annotated data.[17]

Discussion and future work

A promising approach to the design of more appropriate systems for e-disclosure is to design a system using the terms or concepts in which the investigators understand the subject-matter of the investigation. The process in conducting an e-disclosure exercise is one of translating large amounts of unstructured data into representations that are structured by using the terms that lawyers use. The transitions represented in figure 1 can be seen as a process of filtering and abstracting information into these terms.

The approach illustrated in this paper involves identifying how lawyers think about a case and how they want to identify the data. By providing an analysis of the requirements of lawyers and the way these develop, the use of 'frames' can provide a foundation for reasoning about the design in terms of the typical cognitive paths used by lawyers when preparing a case.

If systems can be configured around the concepts that lawyers apply to data, then they are likely to provide a better platform upon which investigators can apply their own expertise in shifting through large volumes of data, and allow them to work to a higher conceptual level.[18] The ideal is that investigators can pursue investigations with fewer interruptions imposed by the constraints of the systems that they use. By identifying documents that are relevant to emerging concepts of the investigation, there is an opportunity to reduce the very high costs of reviewing documents.

These ideas will be explored further in future work. The authors are about to embark on a further case study. Of significant interest will be the way in which lawyers think about the problem as expressed through the language they use and the ways in which they choose to organize information during the process of analyzing it.

A laboratory study is also planned, in which lay participants will perform a mock investigation using a subset of the Enron e-mail collection. This study will involve the presentation of a collection of documents with visual indexes based around different kinds of document aggregation, including e-mail threads, semantic clusters and references to events. The aim will be to understand the value provided by different kinds of methods that can be used to illustrate the organization of documents on screen in the process of identifying cues, then elaborating and validating the related conceptual frames.

Simon Attfield is a Senior Research Fellow at University College London Interaction Centre. He holds a PhD in Computer Science (UCL) and an MSc in Psychology (Sussex). He is currently working on the EPSRC funded Making Sense of Information project, researching how professionals make sense of information in the processes of their work.

http://web4.cs.ucl.ac.uk/uclic/annb/MaSI.html
s.attfield@cs.ucl.ac.uk

Ann Blandford is Professor of Human-Computer Interaction, Director of University College London Interaction Centre, and Principal Investigator on the Making Sense of Information project. She completed her PhD in Artificial Intelligence at the Open University. She focuses on how users experience digital libraries, and how people work with and make sense of information.

a.blandford@cs.ucl.ac.uk

[15] *Robert Gaizauskas and Yorick Wilks, ' Information extraction: Beyond Document Retrieval' , Journal of Documentation, 54(1), 1998, pp 70-105.*

[16] *John Stasko, Carsten Gorg and Zhicheng Liu, ' Sensemaking Across Text Documents: Jigsaw' , Sensemaking Workshop, ACM SIGCHI Conference, 2008, http://dmrussell.googlepages.com/sensemakingworkshoppapers.*

[17] *Inderjeet Mani and George Wilson, ' Robust Temporal Processing of News' , in Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, 2000, pp 69-76.*

[18] *Jens Rasmussen, Annelise Mark Pejtersen and L. P. Goodstein, Cognitive Systems Engineering, (John Wiley & Sons, Inc. New York, 1994).*