Looking for Fraud in Digital Footprints: Sensemaking with Chronologies in a Large Corporate Investigation

Simon Attfield Ann Blandford
s.attfield@cs.ucl.ac.uk a.blandford@ucl.ac.uk

UCL Interaction Centre
Remax House, 31/32 Alfred Place
London, WC1E
http://www.uclic.ucl.ac.uk/

ABSTRACT
During extended sensemaking tasks people typically create external representations that integrate information and support their thinking. Understanding the variety, role and use of these is important for understanding sensemaking and how to support it effectively. We report a case-study of a large, document-based fraud investigation undertaken by a law firm. We focus on the construction and use of integrated representations in the form of chronologies. We show how these supported conjecture recording, focussing on time-periods, identifying gaps, identifying connections and reviewing interpretations. We use our findings to highlight limitations of a previous analysis of representations in sensemaking which regards this as schema definition and population. The findings also argue for search tools designed to identify date references in documents, for the support of ad-hoc event selections, and the support of linking between integrating representations and source documents.

Categories and Subject Descriptors

General Terms

H.1.2 [User/Machine Systems]: Human Factors, Human Information Processing.

Keywords

sensemaking, information interaction, investigation, legal, case-study, chronology

1. INTRODUCTION
Sensemaking is the process through which people use information to construct, maintain and reconstruct interpretations of the world. Its significance for human computer interaction arises where technology can, or does, play a role in supporting interactions with information from which sense needs to be made.

Whilst Sensemaking as a research focus is attracting increasing attention in HCI [see for example, 3, 4, 6, 9], there is a lack of empirical field studies exploring the rich variety of sensemaking tasks [5]. During extended sensemaking people typically create external representations to integrate information and support their thinking [7]. Understanding these, their variety, and how they are created and used is essential for understanding sensemaking and how to support it more effectively.

With this in mind, we report a case-study of a document-based fraud investigation by a large, corporate law firm. The investigation occupied 30 lawyers and forensic accountants for four months. A central representation used by the investigators for integrating and reviewing information were detailed chronologies of events constructed on the basis of evidential documents. These were created using Microsoft Excel spreadsheets, in some cases with additional functionality. We report on how the chronologies were constructed, how they supported the investigators' thinking, and limitations that they presented.

In the next section we briefly review the idea of sensemaking as it is being understood in HCI. In section 3, we describe the case-study method, and in section 4 report our findings. In section 5 we discuss implications of our findings for how we understand the role of representations in sensemaking and in relation to system design.

2. SENSEMAKING

Things make sense when a relationship holds between information and an interpretation (or mental model [5]). It is when such a correspondence fails (or a plausible interpretation cannot be constructed) that something does not make sense. Sensemaking is often described as a process of finding structure by placing information into a framework [6, 7, 8, 11].

Within HCI, sensemaking research has emerged with a focus on understanding the processes of making sense of large bodies of information and the role of representations as sensemaking resources. For example, Russell *et al* [9] present a model based on a case study of a team designing a training course using a hypermedia knowledge structuring tool. They represent sensemaking as an iterative process (the 'learning loop complex') in which schemas (in this case related to entities within the course content domain) are created to capture information, information is sought to populate schema instantiations, and 'residue' (ill-fitting or missing data and unused representations) prompts adjustment to schemas.

Card and Pirolli [4] and also Bodnar [3], present a similarly iterative model which emphasises the interplay between data-driven (bottom-up) and theory-driven (top-down) processes.

Qu and Furnas [8] extended on Russell *et al's* model through a study of participants searching the Web and gathering content for creating presentations. They reported the triggers used for deciding how to structure information and found greater integration between searching for representations and encoding data than expressed in Russell *et al's* model.

Notwithstanding studies such as these, there are few empirical studies which examine sensemaking as it relates to HCI *i.e.* sensemaking with technology. We extend the current view through a case-study of a large sensemaking task in a legal domain. We pay attention to the construction and use of chronologies (primary integrating representation for the investigators), and how these supported the investigation.

3. METHOD

Data were gathered through 10 in-depth interviews with 9 lawyers who had worked on a large fraud investigation. Interviews varied from 45 minutes to 1hr 40. Participants included a Senior Partner, a Junior Partner, a Senior Associate, 3 Associate Solicitors, 2 Trainee Lawyers and a Paralegal. Key artifacts were gathered including a sample chronology.

The interviews were recorded and transcribed. Analysis included creating summary narratives and models, open coding for thematic analysis, selective coding and constant comparison between analysis products and raw data [10]. Analysis products were also used as prompts and to support member checking.

A final in-depth review of this paper was conducted by the Senior Associate (with responsibility for day-to-day running of the investigation) to check accuracy of the account.

Throughout the study, particular attention was paid to understanding activities surrounding the creation and use of external representations, and the way in which these were used as a source of reflection on the investigated domain.

4. FINDINGS

4.1 Setting the scene

The objective of the investigation was to discover whether fraud had occurred within a company, and if so who had been complicit, so that the company could subsequently be 'cleansed'. This was focused by some allegations relating to specific types of business activity. The investigators' raw data consisted primarily of documents (mostly electronic) recovered from computers used by the company. In total, information was recovered from around 500 locations (including email servers) and was equivalent in size to about 8.5 million novels. Other information sources included telephone records and interviews with company staff. As documents were recovered, they were added to a server that supported full-text search.

Given the allegations, and following an initial review of a small collection of key documents, the investigators initially defined a set of investigation issues. They then divided into a series of teams, each assigned to one or two issues. Queries were constructed to retrieve documents that might be relevant to each of the teams' interests, and a team member read and coded each document for relevance. Highly relevant documents were selected as raw material for constructing the chronologies.

As the investigation progressed new sub-issues came to light and different theories were formed. Some queries were dropped, others developed, and queries that were initially broad became richer and more specific. As the issues became refined into sub-issues, so the relevance coding scheme was extended and refined.

4.2 Constructing the chronologies

Each sub-team generated event chronologies for its particular issue(s). An anonymised entry is shown in figure 1. This shows the date and time of an event, a summary description, a list of people involved in the event, and reference to the supporting evidential document(s).

A document could either pre-date or post-date an event to which it referred. Where a document pre-dated an event (e.g. an email discussing a future meeting) it would not necessarily be concluded that the event took place, and so it would be recorded with a qualification.

> Participant 11: […] you would probably put something like, you know "possible meeting" you know "see document so-and-so".

Date	Time	Event/Document	People Involved/ Author/Recipient	Evidence / File Reference
8th Nov	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 Nov. {person I} to pick up {person B} at Airport	{person I}, {person B} and {person H} in {location C}	Email between {person I}, {person I}, {person H} and {person F}/Doc ID 169246

Figure 1. An anonymised extract from one of the chronologies

Possibly at a later time, corroborating evidence would be sought (or simply found) in the form of evidence dated after the event:

> Participant 11: And then you might find later […][an] expense claim or you know, you find out actually telephone conversations happened instead […] And we had quite a bit of that.

Each team generated one or more chronologies relating to their investigation issues. As these were developed, they were integrated into a single master chronology, also created in Excel. The master chronology repeated the format of the issue chronologies, with the addition of codes indicating the relevance of each entry to the many issues. Functionality was added to this chronology such that code sub-sets could be selected (using drop-down lists) and the chronology collapsed to show only those entries relevant to the selected codes.

4.3 Chronologies supporting reflection

4.3.1 Structuring the information space

Given the allegations, the investigators recognized that there were particular periods within the contract lifecycles of the company under investigation that they were particularly concerned about. Given the size of the information universe, focusing attention on these periods was particularly important:

> Participant 6: So some time-periods where it was absolutely critical to know… because you're following this through forensically trying to figure out what's going on… it's absolutely critical to know minute-by-minute the exact chain of events.

In the first instance, the investigators were unsure where these periods would have occurred, or even the identities of the contracts in question. As the chronologies began to take shape, so the contracts and their key events began to emerge. Some documents proved more helpful in this regard than others:

> Participant 4: it may be someone's email saying […] these are key milestone dates […] and on this day I'm planning on being in [overseas city].

However, the investigators did not have retrieval tools that could specifically target documents that would provide this kind of overview.

4.3.2 Spotting gaps

Ultimately, the investigators were able to map out key contract events in broad terms and so define areas of particular interest. This then guided more focused document retrieval using date-delimited queries.

The role of chronologies was to maintain representations of what was known to be the case. By reviewing them as (ideally) consistent, coherent and interconnected narratives, the investigators identified explanatory gaps.

> Participant 4: Well you're kind of thinking why on earth in the middle of a really hectic […] schedule is this guy sending emails saying 'I've got to fly to [country] tomorrow but I'm only going for the day and then I'll be back'

> Participant 4: seeing receipts for bank transfers between people and you kind of think […] why is this money being transferred […] and then you develop some theories for what the sum of money may actually represent

These gaps in explanation lead to theorising, particularly where sinister interpretations were possible. These then prompted theories and still more focused questions.

4.3.3 Seeing connections

The issues underlying each of the teams' chronologies were often interconnected, but in ways that the investigators couldn't anticipate a priori. And so events in one chronology could relate to events in another. Since the master chronology integrated all the findings, the additional view filtering functionality provided a valuable tool for eliminating entries so that events from a limited set of issues could be aligned and considered together.

> Participant 4: […] you could drill down into this chronology and say right we want all documents involving this issue in this time period and it allowed you … let's say if you looked at a snapshot of a period of a week, having assimilated everyone's chronologies together it then allowed you to make links between the different issues that people would never have thought of individually.

> Participant 6: I think the biggest advantage of the collapsible chronology is a fairly obvious one where generally speaking […] you just want to be able to home in on five entries on a certain date, or on events involving two or three people […] you just want the bare minimum that you need to get the answer.

The master chronology filtering was, however, limited to different combinations defined at the unit level of issue codes; and these units were relatively large. Given local areas of density, restricting the view to issues subsets meant that connected events could still be distributed widely and considering relationships could present difficulties.

4.3.4 Reviewing Interpretations

Each chronology entry represented an interpretation of the referenced evidence (see figure 1). During the investigation the (manual) references were used frequently to access source evidence for review. For example, an investigator wishing to record an event could find it already noted. They would then retrieve the hard-copy evidence and judge whether anything new needed to be added.

Reviewing chronology entries also led investigators to identify apparent inconsistencies or ambiguities, prompting them to review source evidence:

> Participant 5: Sometimes you'd go […] "My god, that's a typo, that can't be right because we know from this, that's a much better source of evidence, that that person wasn't in the country". So you'd constantly be revising and reviewing the material.

Finally, the investigators held frequent review meetings. Here they would discuss interpretations, for which access to source evidence was very important.

> Participant 4: twice a day, mid morning, mid afternoon, I would sit down with the six other people in my review team and we would each talk about […] the things we found and then start making various theories
>
> Researcher: How important [..] was it to have the documents that provided the evidence?
>
> Participant 4: Very much […] I know this is something I'm going to want to talk about […] so I print this thing out, physically have it to hand.

5. DISCUSSION

5.1.1 A theoretical issue: where are the schemas?

According to Russell et al's learning loop complex, the sensemaker typically creates schemas to accommodate relevant information, identifies information to instantiate those schemas, and then revises the schemas on the discovery of residue [9]. Whilst this model describes Russell et al's case-study (i.e. designing a training course using a hypermedia knowledge structuring tool), it is not applicable to the study reported here. The investigators did not instantiate predetermined event sequences in the chronologies.

This is not to say that the investigators would not have drawn on internal schemas (such as scripts) when interpreting information. During sensemaking people necessarily use background knowledge to form interpretations of data. The point is that such structures did not explicitly constraint the external representation. The process of making sense was not one of instantiating a finite set of preformed representational structures. The role of external representations in sensemaking is perhaps better understood as a means for recording information and as a source of reflection for prompting further enquiry.

5.1.2 Design recommendations

The current study also provides insights about the design of systems for supporting sensemaking from episodic data, such as in legal investigations. Documents that refer to particular dates are an important resource for validating conjectured events, and where multiple dates are mentioned, these helped in mapping out time-scales and subsequent focusing. Consequently, information extraction tools that interpret date references within documents, or identify documents that refer to multiple dates can improve the efficiency of these activities.

The chronologies were key representations for the investigators, and the ability to visually filter entries in the master chronology provided an effective means for considering causal connections between events. This filtering, however, was limited to combinations defined at the unit level of issue codes, and these units were relatively large. The capability of iteratively selecting records and setting them aside, perhaps in addition to code filtering, would provide greater flexibility for exploring and discussing different possibilities.

A final point is that, during the investigation, document references within event entries allowed the chronologies to act as indexes supporting the re-retrieval of raw evidence. These links were particularly important where interpretations needed to be reviewed. However, these links were not automated. More efficient access would be supported if the source documents could be accessed directly for the summary representation.

To generalize this point, sensemaking tasks are by their nature ill-structured and uncertain, and source documents can be referred to multiple times throughout. It has been shown, for example, that an advantage exists for journalists in being able to automatically maintain links between snippets gathered from digital cuttings services and source documents during news research and writing tasks [1, 2]. We expect this finding to generalize to many scenarios where users create integrating representations to help them make sense of information.

6. Conclusions

Through this case-study we have demonstrated how constructing an external representations for sensemaking is not always a question of schema instantiation. Whilst Russell *et al's* study [9] provides valuable insights, in sensemaking terms we can regard it as a particular kind of example and the current study as another.

Sensemaking can be complex and resource-intensive. The investigation we have reported was extremely costly in terms of time and effort. Our findings suggest some ways in which systems could assist users in reducing the resource overheads in similar tasks. Specifically, we argue that advantage can be found by:

- Supporting the retrieval of documents that refer to specifiable dates, or a number of dates.
- Supporting the iterative and ad-hoc selection of event sub-sets so that these can be aligned and viewed independently of the main collection.
- Supporting linking between integrating representation and source documents.

7. ACKNOWLEDGMENTS

We would like to thank Freshfields Bruckhaus Deringer for their kind help with this study. The work was funded under EPSRC grant ...

8. REFERENCES

[1] Russell, D.M., Stefik, M.J., Pirolli, P., and Card, S.K. The Cost Structure of Sensemaking. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '93)* (Amsterdam, The Netherlands, May, 1993). ACM Press, New York, NY, 1993, 268-276.

[2] Card, S. and Pirolli, P. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. In *Proceedings of 2005 International Conference on Intelligence Analysis* (McLean, VA, USA, May, 2005). http://cryptome.org/intel-2005.htm.

[3] Bodnar, J.W. Making Sense of Massive Data by Hypothesis Testing. In *Proceedings of 2005 International Conference on Intelligence Analysis* (McLean, VA, USA, May, 2005). http://cryptome.org/intel-2005.htm.

[4] Klein, G., Moon, B. and Hoffman, R.R. Making Sense of Sensemaking 2: A Macrocognitive Model. *IEEE Intelligent Systems, 21*, 5, (2006), 88-92.

[5] Qu, Y. A Sensemaking-Supporting Information Gathering System. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '03)* (Ft. Lauderdale, Florida, USA, April, 2003). ACM Press, New York, NY, 2003, 906-907.

[6] Klein, G., Moon, B. and Hoffman, R.R. Making Sense of Sensemaking 1: Alternative Perspectives. *IEEE Intelligent Systems, 21*, 4, (2006), 70-73

[7] Qu, Y., and Furnas, G.W. Sources of Structure In Sensemaking. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '05)*(Portland, OR, USA, April, 2005). ACM Press, New York, NY, 2005, 1989-1992.

[8] Weick, K. *Sensemaking in Organizations*. Sage Publications, Thousand Oaks California, 1995.

[9] Strauss, A. and Corbin, J. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. 2nd ed.* Sage, London (1998).

[10] Attfield, S. Information seeking, gathering and review: Journalism as a case study for the design of search and authoring systems. PhD Thesis, University of London, UK, 2005.