Chapter 3

VISUALIZING INFORMATION IN DIGITAL FORENSICS

Grant Osborne, Hannah Thinyane and Jill Slay

Abstract The evolution of modern electronic devices is outpacing the scalability and effectiveness of the tools used to analyze digital evidence recovered from them. Indeed, current digital forensic techniques and tools are unable to handle large datasets in an efficient manner. As a result, the time and effort required to conduct digital forensic investigations are increasing. This paper describes a promising digital forensic visualization framework that displays digital evidence in a simple and intuitive manner, enhancing decision making and facilitating the explanation of phenomena in evidentiary data.

Keywords: Digital forensic workflow, digital evidence, visualization

1. Introduction

Advances in modern electronic devices are outpacing the ability of digital forensic tools to analyze digital evidence [1, 3]. The two key challenges facing digital forensic investigations are the complexity and volume of digital evidence. The complexity arises from the heterogeneous and idiosyncratic nature of digital evidence; evidentiary data is spread across multiple devices, each with its unique mechanisms for storing and presenting data. Meanwhile, the volume of digital evidence continues to grow as storage becomes cheaper and increasingly massive, and faster processors and high-bandwidth connectivity enable modern electronic devices to be used almost anywhere all the time.

This paper attempts to address the complexity and volume challenges by applying information visualization techniques. Specifically, the paper describes the Explore, Investigate and Correlate (EPIC) process that enhances digital forensic investigations by integrating information visualization techniques into existing digital forensic investigation workflows.

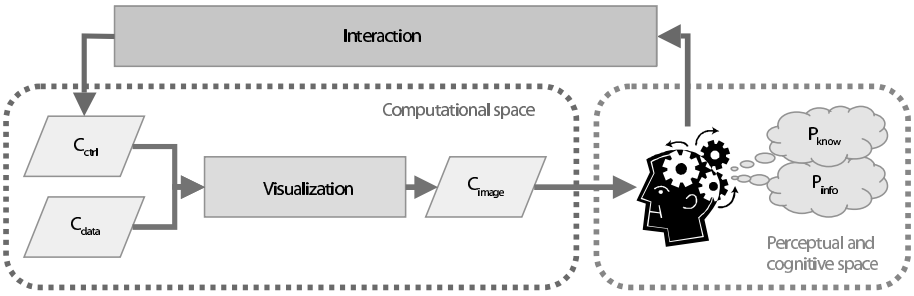

Figure 1. Visualization as a search process [4].

The EPIC process builds on the "visual form" components of the visualization reference model developed by Card, *et al.* [2] (Figure 1). The purpose of visualization is to output an image to a user in a manner that facilitates the understanding of the underlying information. The inputs to a visualization are the visual data to be presented and the control data provided by human interaction. Two "spaces" are involved in visualization as described in the search process model proposed by Chen, *et al.* [4]. The first is the "computational space," where the visualization is updated, controlled and new views are generated. The second is the "perceptual and cognitive space," where the user views the image and gains information and knowledge about the data being visualized and searched. Outside these spaces is the human interaction with the visualization system, which controls the visualization and the generation of updated views. The resulting EPIC process visualization framework can display digital evidence in a simple and intuitive manner, enhancing decision making and facilitating the explanation of phenomena in evidentiary data.

2. EPIC Process

The EPIC process has seven goals: (i) make the evidence visible; (ii) reduce the relative size of the evidence; (iii) provide high-level overviews of the evidence; (iv) aid in the presentation of the events and relationships in the evidence; (v) provide explanations of the origin and significance of the evidence; (vi) provide support to identify items of probative value; and (vii) facilitate the presentation of the findings to other investigators or in court.

To meet these goals, the EPIC process shown in Figure 2 combines a graphical representation of digital evidence with a search process. The EPIC process contains a computational space that supports multiple visualization techniques: updating, deleting and adding data, and up-

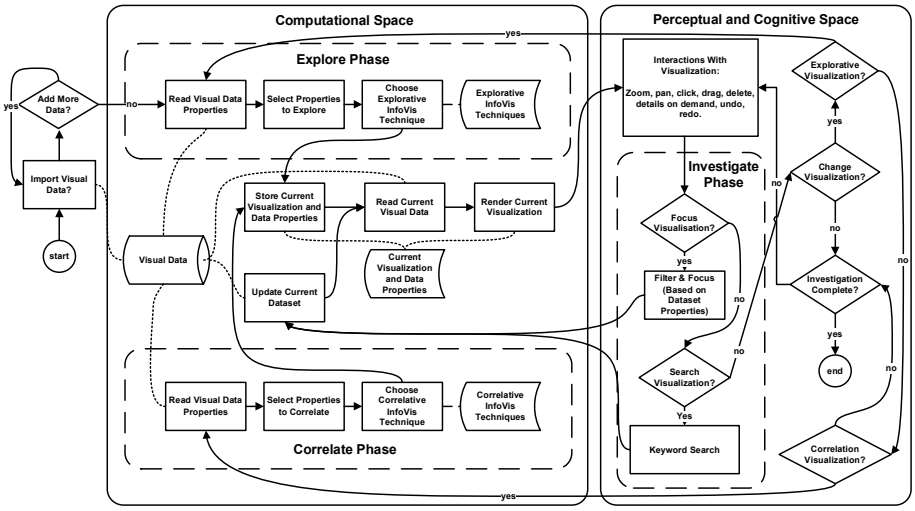

Figure 2. EPIC process model.

dating the current visualization technique in real time. The two main phases in the computational space are explore and correlate. These two phases focus on selecting data properties, mapping them to appropriate visualization techniques and displaying the views to the user.

The EPIC process guides an investigator through a set of tasks during the examination of the available digital evidence in visual form. The tasks aid an investigator in selecting an evidence property to visualize and a visualization technique for presentation. Additionally, the EPIC process includes a common set of digital forensic tools to enable the investigator to investigate, focus and filter the digital evidence within the visual data.

3. User Study

Since the application of information visualization techniques to digital forensic investigations is a new concept, a user study was designed to investigate the ideas underlying the EPIC process. The study sought to examine if the EPIC process improved the analysis and presentation of large quantities of digital evidence from distributed sources.

The EPIC process was compared with two other approaches. The first is a textual display visualization approach as used in industry tools such as EnCase and FTK. The second is a simple visualization approach, which incorporates representation and interaction techniques to iden-

tify if process-driven information visualization provides advantages over user-driven visualization.

The user study evaluated two usability goals. The first is the effectiveness of the EPIC process to aid the analysis and presentation of digital evidence compared with the textual display and simple visualization approaches. The second is whether the EPIC process reduces problems associated with the analysis and presentation of large quantities of digital evidence.

To evaluate these goals, the user study collected data from an information gathering component in the information visualization system along with participant feedback. The information gathering component collected participant performance metrics when the approaches were used. Two metrics for each participant were recorded. The first was the time taken to answer a question relating to a fictitious scenario presented in the user study. The second was the number of attempts made by a participant to determine the correct answer to the question.

Several questionnaires were used to obtain participant feedback for the user evaluation study. A questionnaire was presented pre-study and post-study for each of the approaches used by the participants. Each questionnaire focused on participant opinions about various characteristics of the approaches, and participant experience in digital forensics and using information visualization tools.

The usability questions were derived from the IBM System Usability Satisfaction Questionnaire [6]. The questions focused on participant opinions on the usability of an approach in achieving the tasks put forward in the user study. The NASA Task Load Index (NASA-TLX) Questionnaire [5] provided the basis for the task load index questions. The questions focused on the mental demands imposed by the visualization approaches and the effort undertaken by the participants to achieve their goals. The ranking of the three approaches was based on the participants' opinion of their effectiveness at completing tasks, familiarity compared with the participants' previous experience with other analysis tools, and ease of learning if the approaches were to be used in the participants' daily workflow. The familiarity and ease of learning rankings reflect the participants' overall opinions of the learnability of the approaches.

3.1 Procedure

The evaluation phase of the user study presented each participant with event-based data for use with each of the three approaches: textual display visualization, simple visualization and EPIC process visualiza-

tion. The data was created to represent a fictitious criminal scenario with a series of events. Each participant was required to answer two questions related to the scenario by interacting and working with each approach. In total, six unique questions were asked of each participant.

The evaluation phase also required each participant to complete usability and task load index questionnaires for each approach that was used. The IBM System Usability Satisfaction Questionnaire required participants to rank each approach from 1 to 5 for ten usability related questions (higher score is better). The NASA-TLX Questionnaire also required participants to rank each approach from 1 to 5. However, since this questionnaire focused on the mental demand, a lower score is better.

The final component of the user study was the post-intervention questionnaire. This questionnaire attempted to capture the participants' overall thoughts about the three approaches. A combination of drop-down boxes and textual input was used to obtain ranking data. In particular, the participants were required to rank the approaches from highest to lowest with respect to:

- **Effectiveness:** How well an approach can answer the questions about the scenario.
- **Familiarity:** How familiar is an approach.
- **Ease of Learning:** How easy it is to learn an approach.

In addition to the rankings of the approaches, the participants were required to provide reasons for their rankings.

3.2 Dataset

The software used for the user study presented information as a series of discrete "events" between a source entity and a target entity. The study included the following event types: email, Short Message Service (SMS), Multimedia Messaging Service (MMS), phone call and website visit. These events were chosen because they provide a good representation of the types of events (other than images) that investigators work with when piecing together a case from digital evidence sources.

Since a publicly available dataset suitable for this evaluation did not exist, a dataset was created using the event types with multiple digital evidence sources. In particular, the dataset was created to represent a mock criminal case with a realistic number of sources and key evidentiary events. The dataset contained more than 100 unique events with approximately 30 of the events directly related to the fictitious criminal scenario. The other events were designed to be noise consistent with a

digital evidence dataset. The events were distributed among nine digital evidence sources, namely computers and mobile phones. Seventeen unique entities (persons and websites) were added to the dataset to serve as the sources or targets of events in the dataset. The number of sources used was based on the numbers suggested by Turnbull, *et al.* [7]. Other studies of visualization techniques in digital forensics [1] have created similar datasets for evaluation.

3.3 Software Environment

The participants were required to interact with a custom software environment that was created for this research. The software facilitated the creation of entities, events and digital evidence sources in the fictitious criminal scenario. Also, it enabled the data to be preloaded into a database for presentation to the study participants.

The software recorded participant performance when using the three approaches to analyze digital evidence. A participant was asked a question regarding the scenario, which was broken down into components for the participant to answer. The participant was then required to use the given visualization approach to find one or more events within the dataset that provided the information required to answer the various components of the question. If the participant entered an incorrect answer, the system recorded this fact and gave visual feedback about the incorrect components. The software recorded the time taken to answer each question and the number of incorrect attempts for each component.

In addition to the evaluation phase questions, the software presented the participants with questions to obtain feedback about the visualization approaches. A similar interface was used for the pre-intervention and post-intervention questionnaires.

3.4 Visualization Approaches

The three visualization approaches evaluated were: (i) textual display visualization; (ii) simple visualization; and (iii) EPIC process visualization.

The textual display visualization approach shown in Figure 3 is similar to those provided by industry tools such as EnCase and FTK. The approach incorporates filtering and searching functionality to help focus the dataset and filter unimportant information in real time based on constraints imposed by the investigator. The dataset filters shown in Figure 3 include entity, event type, keyword and date range. The colors for the event type filters in the filter panel correspond to those used in the textual display visualization approach.

Figure 3. Textual display visualization.

The simple visualization approach was applied to social network visualization, specifically for highlighting relationships and behavioral structures. The approach incorporates common visualization techniques such as zooming, panning and acquiring details on demand about a particular person and his/her relationships in the graph.

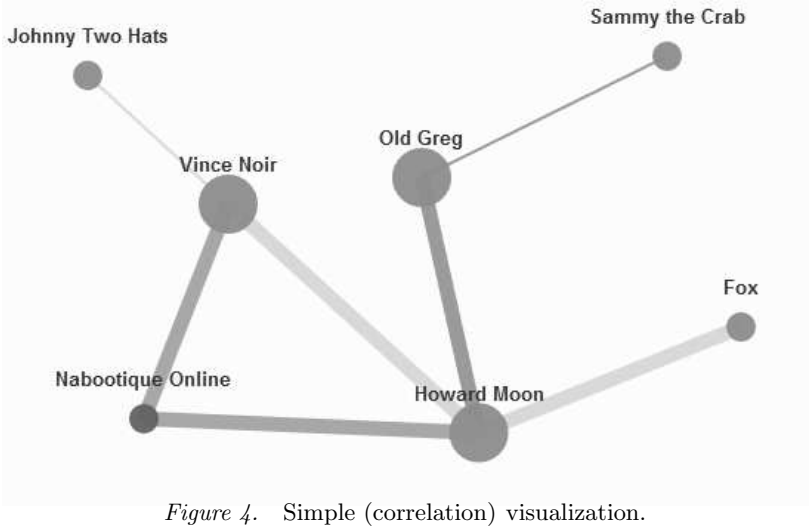

Figure 4. Simple (correlation) visualization.

Figure 4 shows the simple visualization approach. The circles in the visualization represent people or websites. The size of an entity represents the number of events for which it is the source. The thickness of the event line is based on the number of events that link the two entities. The event colors used in the textual display visualization approach are also used in this visualization approach to enhance participant understanding.

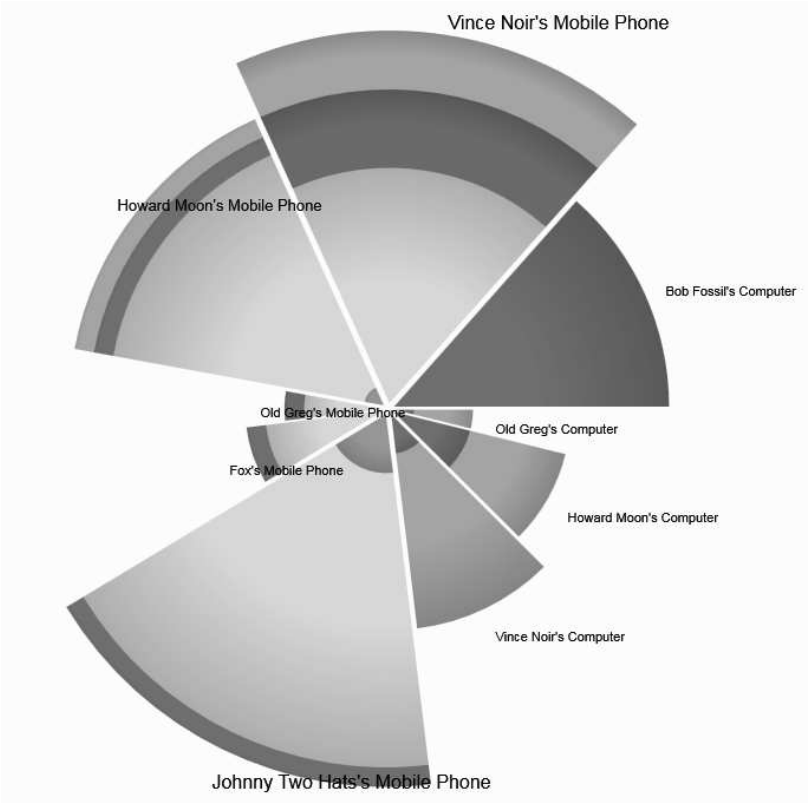

Figure 5. Explore visualization.

The third approach, EPIC process visualization, incorporates all the critical tasks, phases and visualization techniques defined by the EPIC process. The visualization approach provides the filtering and searching functionality of industry standard tools, and corresponds to an implementation of the investigate phase of the EPIC process. The results of the investigate phase are displayed in real time by the correlation and explore visualizations in Figures 4 and 5, respectively.

The explore visualization in Figure 5 is a pie chart visualization that shows all the non-filtered events as a high level overview. Each piece of the pie represents a digital evidence source, with the layers within the piece representing the different event types and their relative amounts on the source. The pieces of the pie are scaled to give a sense of the sources that contain the most information within the current filter set. In conjunction with the dataset filtering techniques (Figure 3), participants were able to interact with all the visualizations to perform the traditional digital forensic investigation and visual data representation workflows.

3.5 Data Collection

The independent variable in the study was the order in which the visualizations were analyzed (and the questions asked). The controlled variable was the two questions asked per evaluation phase. The dependent variables that measured participant usability of each visualization were the time taken to answer the questions and the accuracy (correctness) of the answers.

The time taken to answer questions was measured in milliseconds by the study software. The clock started when the question was opened and ended when the participant entered the correct answer. The accuracy of an answer to a question was measured by computing the inverse of the number of errors made when answering the question. Note that an error was defined as occurring when a participant did not answer all the components of a question correctly.

The questionnaires also provided a secondary dependent variable, namely participant feedback regarding the usability and task load of each visualization approach. Usability and task load, collectively referred to as satisfaction metrics, were collected by the user study software. The usability score was calculated as the mean of the participant responses ranking the effectiveness, comfort, error recovery, information display and productivity of a visualization approach from 1 (strongly disagree) to 5 (strongly agree). The task load score was calculated as the mean of the user responses from 1 (very low) to 5 (very high) for the questions regarding mental workload, success, and security and stress felt when using the visualization approach.

The third measure in the user study was the participant ranking of the three approaches. The ranking was captured using the post-intervention questionnaire, which required participants to rank each approach from 1 (best) to 3 (worst) on several factors, including effectiveness at completing the questions, familiarity compared with existing analysis tools, and ease of learning the approach if incorporated into the workflow for an extended period of time.

Each participant was also required to provide the reasons for assigning the rankings, which provide a qualitative component for the measure. Because performance measurements (time taken and accuracy) were automated in the user study software, the measures exhibited strong inter-observer reliability, test-retest reliability and internal consistency reliability. The satisfaction metrics (usability and task load) and approach rankings required participant input. This created uncertainty in the reliability of participant responses as well as in the reliability and validity of the questions. To obtain reliable responses, the participants were

given as much time as needed to complete the questionnaires. To ensure reliability and validity, the usability questions were derived from the IBM System Usability Satisfaction Questionnaire [6] and the task load questions were derived from the NASA-TLX Questionnaire [5]. Since picking a "best approach" had the potential to impose additional stress on the participants, the comfort level of the participants was enhanced by providing them with time and space to provide constructive criticism regarding the three approaches.

3.6 Study Hypothesis

The main hypothesis in the user study was that the EPIC process visualization approach would yield performance and satisfaction measures that were higher on the average compared with the textual display and simple visualization approaches. It was hypothesized that the majority of the individual measures would favor the EPIC process visualization approach. However, it was hypothesized that the time taken component of the performance measure would favor the textual display visualization approach because of participant familiarity with the approach in industry tools. Thus, the time taken to complete tasks with the textual display visualization approach was expected to be lower than that for the other approaches.

3.7 Study Participants

The participants in the study included nine digital forensics experts from South Australia. Due to the small sample size, a statistical analysis of the results was not appropriate. The experts were either investigators or computer analysts, all of whom used industry standard tools very frequently to analyze digital evidence. Since the data complexity and volume challenges discussed in this paper directly affect their job performance, the feedback gathered from the experts can be used to improve visualization techniques.

4. Study Results

Figure 6 shows the performance measures (time taken and accuracy) for the three visualization approaches. The EPIC process visualization approach has the second best time taken measure (7.7 milliseconds) and the best accuracy measure (1.0). As hypothesized above, the textual display visualization approach has the best time taken measure (7.2 milliseconds) due to participant familiarity with the visualization approach as implemented in industry tools. Indeed, 85% of the participants stated

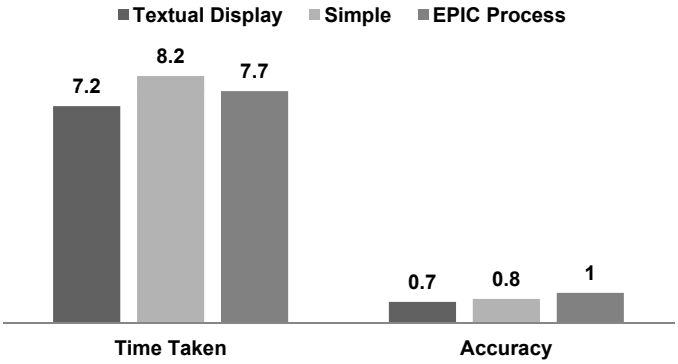

Figure 6. Performance measure results.

they found the textual display visualization approach to be the most familiar.

The performance measures demonstrate that the EPIC process assists investigators in analyzing multiple digital evidence sources to correctly answer questions about digital evidence. Furthermore, the measures highlight that the EPIC process helps achieve quick results, most likely as a result of its inclusion of exploratory visualization techniques as well as common investigative techniques that were familiar to participants.

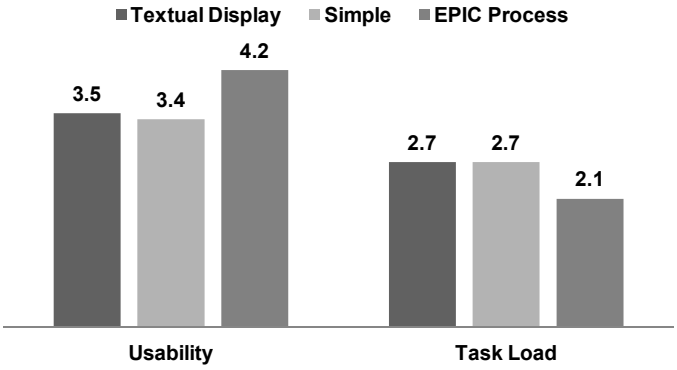

Figure 7. Satisfaction measure results.

Figure 7 shows the satisfaction measures for the three visualization approaches. The usability metric expresses the simplicity, efficiency and performance levels provided by the visualization approaches; a higher score is a better result (maximum value of 5). The task load metric focuses on the mental demand, stress and perceived successes and failures of participants as they used the visualization approaches to answer

questions about the dataset; a lower score is a better result (minimum value of 1).

The EPIC process visualization approach has the highest usability score (4.2). Also, it has the highest task load index (2.1). Thus, the EPIC process visualization approach has the best overall usability score of the three visualization approaches studied. The combination of best practice visualization techniques with familiar investigative technologies enables the EPIC process visualization approach to effectively support the analysis of digital evidence from multiple sources.

The ranking measures provide insight into how the participants rated the three visualization approaches in terms of effectiveness, familiarity and ease of learning. Overall, the highest percentage of participants ranked the EPIC process visualization approach as the most effective. This was primarily due to the reduced complexity of the presented data and the use of a simplified graphical model for dealing with data volume. Thus, the key goals of the EPIC process visualization approach, which are to minimize the impact of data complexity and data volume on digital forensic investigations, appear to have been realized.

However, the majority of the participants ranked the textual display visualization approach as the most familiar approach. This is largely because the approach is implemented in most industry tools. The familiarity ranking of the EPIC process visualization approach could be improved by incorporating techniques used in industry tools. Some participants noted that the EPIC process visualization approach would often oversimplify the data presented to users. Indeed, the participants observed that the approach would have a better familiarity score if more evidence was presented, but in a less complex manner than the textual display visualization approach. Striking the right balance between detail and simplicity is a topic for future research.

5. Conclusions

The EPIC process visualization framework helps display digital evidence in a simple and intuitive manner, enhancing decision making and facilitating the explanation of phenomena in evidentiary data. The user study reveals that the EPIC process visualization approach has the best overall performance and satisfaction measures compared with the textual display visualization and simple visualization approaches. However, textual display visualization was ranked as the most familiar approach, largely because it is implemented in industry tools.

The EPIC process visualization approach can be improved by not oversimplifying the displayed data and making the visualization more

familiar to users. Graphical icons can be used to represent programs and events (e.g., using an envelope for email). Also, content previews can be incorporated similar to the SMS and email previews on iPhones and the Aero Peak content thumbnail in Windows 7.

Future enhancements to the user study include merging the explore and correlate phases of the EPIC process model into an overarching visualization phase. Within this phase, the user would no longer have to switch between multiple visualization categories and could have a constant overview style visualization overlaid on a correlation visualization. This would minimize the cognitive overhead involved in switching from an explore visualization to a correlation visualization and vice versa. Future research will also focus on whether using a constant explorative overview visualization in conjunction with correlation visualization is superior to having a gated progression from simple overview visualization to a more complex correlative visualization.

References

[1] N. Beebe, J. Clark, G. Dietrich, M. Ko and D. Ko, Post-retrieval search hit clustering to improve information retrieval effectiveness: Two digital forensics case studies, *Decision Support Systems*, vol. 51(4), pp. 732–744, 2011.

[2] S. Card, J. MacKinlay and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann, San Francisco, California, 1999.

[3] E. Casey, State of the field: Growth, growth, growth, *Digital Investigation*, vol. 1(4), pp. 241–242, 2004.

[4] M. Chen, D. Ebert, H. Hagen, R. Laramee, R. van Liere, K. Ma, W. Ribarsky, G. Scheuermann and D. Siler, Data, information and knowledge in visualization, *IEEE Computer Graphics and Applications*, vol. 29(1), pp. 12–19, 2009.

[5] S. Hart, NASA-Task Load Index (NASA-TLX): 20 years later, *Proceedings of the Fiftieth Annual Meeting of the Human Factors and Ergonomics Society*, pp. 904–908, 2006.

[6] J. Lewis, IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use, *International Journal of Human-Computer Interaction*, vol. 7(1), pp 57–78, 1995.

[7] B. Turnbull, R. Taylor and B. Blundell, The anatomy of electronic evidence – Quantitative analysis of police e-crime data, *Proceedings of the International Conference on Availability, Reliability and Security*, pp. 143–149, 2009.