2011

Component technologies for e-discovery and prototyping of suit-coping system

Youngsoo Kim
Electronics & Telecommunications Research Institute (ETRI), Korea

Dowon Hong
Electronics & Telecommunications Research Institute (ETRI), Korea

Originally published in the Proceedings of the 9th Australian Digital Forensics Conference, Edith Cowan University, Perth Western Australia, 5th -7th December 2011

COMPONENT TECHNOLOGIES FOR E-DISCOVERY AND PROTOTYPING OF SUIT-COPING SYSTEM

Youngsoo Kim, Dowon Hong
Electronics & Telecommunications Research Institute (ETRI), Korea
blitzkrieg@etri.re.kr, dwhong@etri.re.kr

Abstract

As ESI (Electronically Stored Information) is included in extent of evidence that become discovery's target in FRCP(Federal Rules of Civil Procedure) taken effect on December 1, 2006, enterprises been always vexing in several litigations need to adapt systems coping with e-Discovery such as ESI administration or information preservation. In this paper, component technologies for all steps of e-Discovery are described in detail, and as a prototype of preparing system for e-Discovery, agent-based information management and control system being able to manage ESI stored at some computers centrally and respond rapidly on demand, extracting discovery-related data using digital forensic technologies, are introduced. Apart from fundamental searching and analysing functions, this system can detect user's abnormal behaviours, generate forensic images remotely, and have a function of controlling related files.

Keywords

Digital Forensics, E-Discovery, Enterprise Data Management System, Litigation Hold

INTRODUCTION

In 2005, America's large financial investment firm Morgan Stanley lost a case and compensated Revlon Inc. for $ 600 million and firm's image was tarnished. It is the decisive reason to lose a suit that this firm violated court's discovery request all related e-mails should be submitted. In addition, Samsung Electronics lost a suit at patent dispute and paid a $56 million fine to Israel's Mosaid Inc. in 2004. The reason to be defeated is the same. Samsung Electronics deleted some e-mails could be used as important legal evidences on purpose (Volonino et al, 2010).

The FRCP (Federal Rules of Civil Procedure) has a procedure of asking an opposing party to open related evidences and information through discovery (FRCP, 2006). A litigant opens and collects information and evidences to clarify a point at issue of litigation, by legal method out of court in order to prepare trial. By asking each other to open an opposing party's evidences, documents, or witnesses, it can help litigants proceed this lawsuit under the same condition.

Litigants should open all evidences they have by themselves prior to trial and can request the other party or the third party to make public theirs at the same time. The purpose of this requesting procedure for opening evidences is to make clear a point at issue of suit and secure all evidences which might be hidden purposely on trial, and there are a lot of cases that compromise is achieved prior to trial because each party knows about the other party's evidences in detail (Kim et al, 2010). Discovery is made in writing such as a written request, a written answer, or a written protest and all documents need a lawyer's signature. This process is fulfilled between litigants without a court's participation. However, if a dispute occurs which a litigant rejects requests of the other litigant, a court participates in it. If litigants make excessive or expensive discovery requests on purpose, a court can revoke them, conversely, they do not their duty of discovery in good faith, it can imposes mandatory sanctions.

As ESI (Electronically Stored Information) is included in extent of evidence that become discovery's target in the FRCP taken effect on December 1, 2006, terminology named e-Discovery was appeared (FRCP, 2011). The FRCP governs the conduct of civil actions in the federal courts, and until these 2006 amendments, the guidelines mostly ignored questions regarding digital evidence (Gartner, 2008). The changes, which took effect on 1 December 2006, addressed six areas:

- Meetings between adversaries as well as the judge
- What is reasonably accessible for discovery
- Procedures for handling inadvertent loss of privilege

- Electronically stored information
- Production formats
- Accidental loss of electronically stored information

Enterprises been always vexing in several litigations are hurrying to adopt systematic ESI administration and information preservation system to prevent a lawsuit from losing owing to failure in duty of presenting related evidences and to maintain their confidences (Kim et al, 2010; Cohen et al, 2010; DLP, 2011).

In this paper, component technologies for all steps of e-Discovery are described in detail and, as a prototype of preparing system for e-Discovery, agent-based information management and control system being able to manage ESI stored at some computers centrally and respond rapidly on demand, extracting discovery-related data using digital forensic technologies, are introduced. Apart from basic searching and analysing functions, this system can detect user's abnormal behaviours, generate forensic images remotely, and have a function of controlling related files. At first, we introduce EDRM briefly and describe component technologies for e-Discovery step by step. Additionally, we suggest a fundamental prototype for suit-coping system and conclude.

EDRM (E-DISCOVERY REFERENCE MODEL) AND FUNCTIONS

E-Discovery related tools or solutions are designed and made referring EDRM of figure 2. This reference model standardizes proceedings and defines each step's functional specification to effectively follow guidelines and recommendations described in FRCP (EDRM, 2011). EDRM offers general, scalable, and flexible frameworks to develop e-Discovery related products and services and evaluate them. This is used by general standard about e-Discovery that is authorized, because it was developed by co-works of various related organizations.

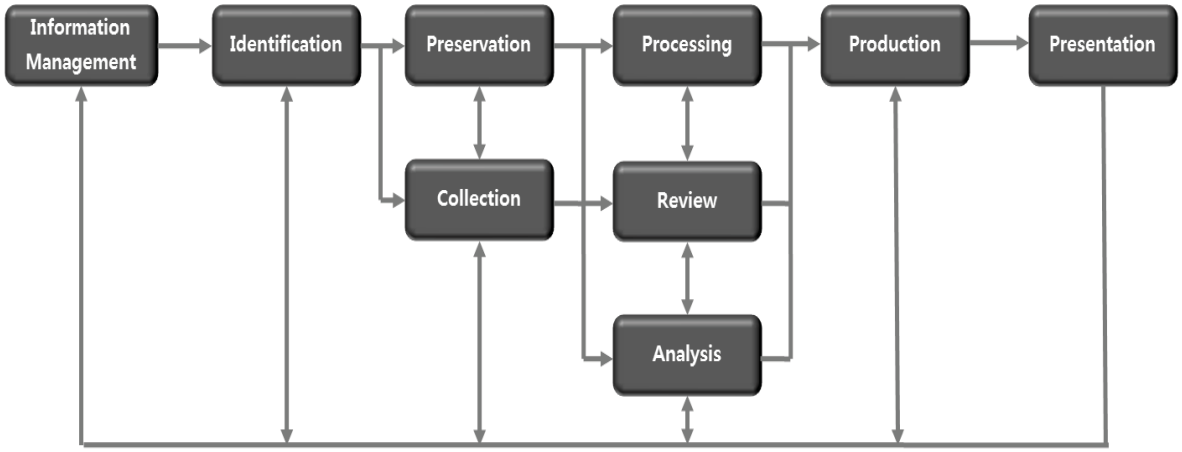

Figure 1: Electronic Discovery Reference Model (EDRM)

Information Management - It manages documents of enterprises or organizations with documents control and preservation policy. **Identification** - As step of deciding a scope of discovery-related documents, it prepares documents can be used in discovery potentially and decides documents should be collected and preserved. **Preservation** - It secures that documents do not change or destroy. **Collection** - It collects ESI from various media such as tapes, drives, portable storage, networks, etc. **Processing** - It filters duplicated or unrelated documents and changes format of ESI to be able to review them more effectively. **Analysis** - As step of making related summaries (Related subjects, persons, or documents) by analysing ESI, it should be done to enhance productivity prior to detailed review step. **Review** - As step of establishing strategy on court, it evaluates collected ESI via relations and privileges and selects sensitive documents. **Production** - It stores ESI to various media and submits it to a court and opposing litigants. **Presentation** - It considers methods that can be seen effectively in trial.

COMPONENT TECHNOLOGIES FOR E-DISCOVERY

We divide technologies for e-Discovery into 8 steps. These steps are assigned referencing EDRM and the Sedona Conference, yet another e-Discovery project (Sedona Conference, 2011). The Sedona Conference is a non-partisan think tank on law and policy. A number of working groups consisting of lawyers, jurists, and consultants that work to provide best practices and commentary on e-Discovery issues address various issues of

ESI and related topics. EDRM focuses on managing ESI and developing related solutions, while the Sedona Conference focuses on responses and workflows of enterprises or organizations in case of lawsuit.

8 steps consist of information management, strategy establishment, collection, preservation, processing, review/analysis, production, and presentation/destruction. Table 1 shows detailed component technologies for e-Discovery.

Table 1: Component Technologies for E-Discovery

E-Discovery Steps	Component Technologies	
Information Management	- Policy Establishment & Adaptation - Policy Compliance Monitoring - Employees' Relationship Definition	- ESI Category Definition - ESI Automatic Classification - ESI Lifecycle Management
Strategy Establishment	- Lawsuit Issues Examination - ESI Search - Early Case Assessment - E-Discovery Planning	- Related ESI Identification - Litigation Hold Execution - Data-map Creation & Management
Collection	- Collecting-method Choice (Integrity) - Identified ESI Backup - Copy Creation (Imaging)	
Preservation	- Policy Establishment & Adaptation - Litigation Hold Management	
Processing	- ESI Evaluation & Data Recovery - Data Format Transformation - Container-File Extraction - Metadata Acquisition - Similarity-based Hash Analysis	- De-Duplication - Near-Duplication Analysis - Target ESI Indexing - Condition-based Filtering
Review & Analysis	- Review Strategies & Planning - Review Format Transformation - Redaction - ESI Re-Search through Review Plan - Visualization of Integrated ESI - Privilege Log Creation	- Tagging or Annotation - Grouping - Reviewing Result Reporting - Context-based Analysis - Relation Analysis between ESI and Suit
Production	- Evidence-Producing Format Analysis - Specific File Format Production - Production Log Creation/Management - Load File Creation - Chain of Custody Log Creation	
Presentation & Destruction	- Evidence Visualization - Unprofessional Report/Diagram Creation - Policy-based ESI Destruction	

Information Management – It is the step of preparing law suits and then is prior to occurrence of a civil suit. Some functions such as Policy Establishment & Adaptation, Policy Compliance Monitoring, Employees' Relationship Definition, ESI Category Definition, ESI Automatic Classification, and ESI Lifecycle Management are required. Document managers of an enterprise set-up various policies for managing information generated, modified or deleted. After setting-up policies, ESI could be managed through the function of ESI lifecycle management. Furthermore, a function of monitoring whether all employees keep them well is required, too. In this step, automatic classification techniques for documents are very useful for finding some evidences could be used at the court later. Therefore, functions of categorizing and classifying ESI automatically using those categories and some classifying algorithms are needed (Yang et al, 1997; Joachims, 1997).

Strategy Establishment – When a law suit starts, litigant parties need this step, first. It requires some functions of Lawsuit Issues Examination, ESI Search, Early Case Assessment, E-Discovery Planning, Related ESI Identification, Litigation Hold Execution, and Data-map Creation & Management. At first, all the people concerned should understand issues of that suit, so they need a searching function of ESI and ECA-related functions. ECA (Early Case Assessment) refers to estimating risk (cost of time and money) to prosecute or defend a legal case (ECA, 2011). ECA lifecycle will typically include the followings: A risk-benefit analysis, information preservation, gathering relevant information, process potentially relevant information for filtering, search term, or data analytics, reuse information in future case, etc. Based on ECA, litigant parties establish e-Discovery plan. They identify related ESI and execute Litigation-hold for preserving them. Creating and managing data-maps can help do above things better.

Collection – After establishing strategy, parties concerned should collect identified information. This step includes Collecting-method Choice, Identified ESI Backups, and Creating Copies (Imaging). They can collect only related data or disks including related data. Integrity should be considered to choose the way of collecting. Usually, parties concerned process or review a copied version of data, not original one, in order to prevent information being changed. The only copying way with integrity is imaging. The copying images from the original data or disks can be used. Furthermore, Identified ESI backups are also one of the prominent functions of this step.

Preservation – This step preserves candidate data which could be used as evidence information at the court and includes functions such as Policy Establishment & adaptation function and Litigation Hold Management. At first, parties concerned setup several policies for preservation like extension types, creation time, recent modification time, employees' name, IP address, MAC address, preservation starting time, preservation period, scope of preservation, preservation method, preservation type, etc. Additionally, parties concerned should monitor whether the Litigation Hold, started at strategy establishment step, is being executed well or not.

Processing – This is a step of processing data to review or analyze. It includes ESI Evaluation & Data Recovery, Data Format Transformation, Container-File Extraction, Metadata Acquisition, Similarity-based Hash Analysis, De-Duplication, Near-Duplication Analysis, Target ESI Indexing, and Condition-based Filtering. Each company stores ESI on several types of media. When the ESI is being created, received, or processed, or when it must be quickly and frequently accessed, it is stored at online storage like hard drives. Some ESI is stored at removable media such as DVDs, CDs, or flash drives. In this case, the files are available in a short period, such as a few minutes. Usually old ESI is stored at offline storage or backup tapes. Offline storage and archives is magnetic tapes or optical disks. It differs from online or removable storages in that the storage media are labelled, organized in shelves or racks, and accessed manually. Backup tapes, commonly using data compression, are sequential access media. The data is not organized for retrieving individual files. Retrieval typically requires restoring contents of the entire tape. In processing step, parties concerned evaluate and recover ESI from above types of media. This step requires also a function of transforming data format in order to review or analysis and extracting function for Container-files (Container Format, 2011). Additionally, it needs to acquire metadata showing file's information and filter well-known files like operating system files not being analysed using hash analysis. De-duplication and near-duplication are essential functions of this step. Through these functions, candidate data to review or analyse can be reduced prominently. Finally, target ESI indexing and condition-based filtering functions are also required in this step. Even though indexing takes long time to complete the index, it is very useful for searching a specific data. Filtering some ESI which mean nothing to reviewers can also reduce a respectable amount of reviewing data.

Review and Analysis – This is a step of extracting evidence data from processed one. Various functions are included such as Review Strategies & Planning, Review Format Transformation, Redaction, ESI Re-Search through Review Plan, Visualization of Integrated ESI, Privilege Log Creation, Tagging or Annotation, Grouping, Reviewing Result Reporting, Context-based Analysis, Relation Analysis between ESI and Suit, etc. After reviewing plans are made, ESI are re-searched through reviewing plans. Parties concerned use redaction, tagging, annotation, or grouping for providing convenience for reviewing and analysis. For high-level analysis, several functions such as visualization of integrated ESI, Context-based analysis, or relation analysis between ESI and suit are used. A function of privilege log analysis is also essential. Confidential conversations and communications that are protected by law from being used as evidence or revealed to others are referred to as privileged. Unless there's an exception, privileged ESI is not discoverable. Therefore, privileged ESI should be handled carefully using this function.

Production – This step transforms reviewed data to specific format files to present to the court. It needs several functions like Evidence-Producing Format Analysis, Specific File Format Production, Production Log Creation/Management, Load File Creation, Chain of Custody Log Creation, etc. After analysing evidence-

producing format, parties concerned select a specific file format and create load files. Functions of loggings like production log or chain of custody log are also needed.

Presentation and Destruction – This is the Final step of e-Discovery. Functions like Evidence Visualization, Unprofessional Report/Diagram Creation, or Policy-based ESI Destruction are needed. To understand presented files in the court well, a visualization function is useful. Presentation report should be made unprofessionally and it is a better way to add several easy diagrams. After presenting, litigant parties destruct all ESI through some policies.

SYSTEM ARCHITECTURE AND MAIN FUNCTIONS

System Architecture

Figure 2 depicts system components and composing blocks. This proposed system has 3 components, server, manager and agents and comprises 6 blocks, EFSB(Enterprise Forensic Server Block), EFMB(Enterprise Forensic Manager Block), EFAB(Enterprise Forensic Agent Block), EFUB(Enterprise Forensic User Block), EFTB(Enterprise Forensic Authentication Block), and EFGB(Enterprise Forensic GUI Block).

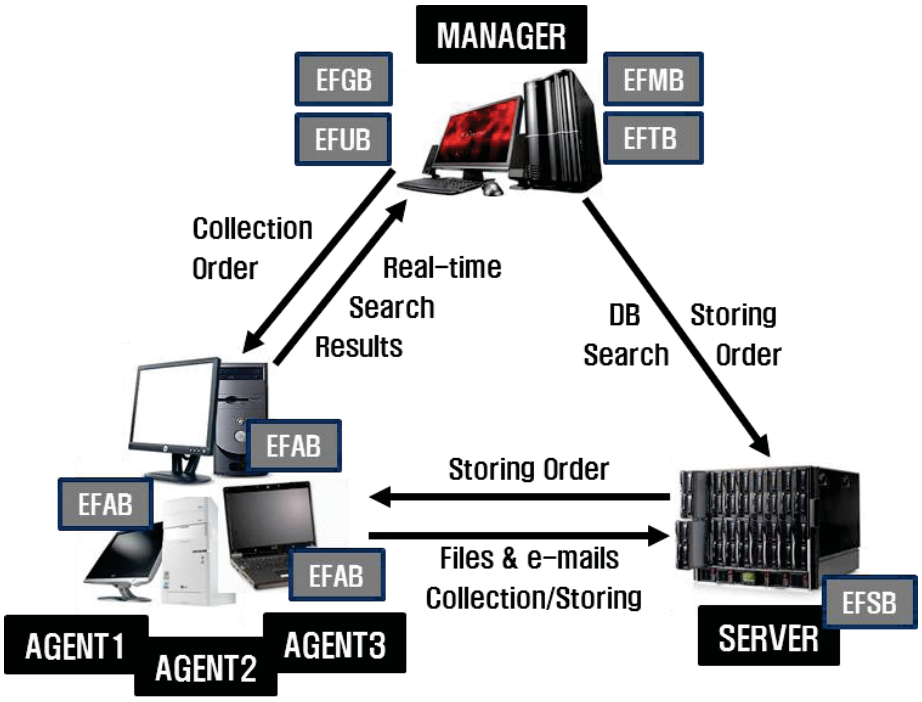

Figure 2: System Components & Composing Blocks

All agent PCs are loaded with EFAB. This block indexes characters of document files at agent PCs and sends some specific files to EFSB in obedience to EFMB's orders. Additionally, it can stop users deleting target files or e-mails in case of Litigation-Hold, monitor files' life-cycle such as generation, modification, or discarding, and generate and transmit remote forensic images. EFSB operating in the server stores contents and attributes of files EFAB extracted in database and manages them. Furthermore, this block stores and administrates file copies and hash values according to EFMB's order.

A manager has 4 blocks, EFMB, EFGB, EFTB, and EFUB. EFMB can search or review an index stored already at EFSB and play functions such as real-time investigation and retrieval, deletion-blocking set-up, file monitoring, forensic image generation, etc. Additionally, this block can preview completed working contents check contents of e-mails. EFGB receives user's order using a keyboard and a mouse and sends it to EFMB and displays operating results on the screen. EFTB, certification block for users or administrators, receives certifying information from EFMB or EFUB and returns a certification result after fulfilling a certification routine. EFUB being empowered by EFTB connects to EFSB and searches already indexed characters or checks transmitted files or e-mails by the authority of reviewer. Main functions that our system provides are as follows.

Real-time File Collection

An agent sets up collection-policies on real-time and collects files and then a manager can check the result directly on real-time. Collecting results are not submitted to the server. A manager selects attributes of files and sends queries to EFAB to have a real-time search for files satisfied with some specific conditions using EFGB. A file's attributes such as keywords being included in contents of files, file metadata (documents' generation time, modified time, recently accessed time, file sizes, file name, file extension, or file owner (user account name)), or file hash values are used in this case. After receiving queries, EFAB searches target files on real-time and then sends results to EFMB. Finally, EFMB sends these results to EFGB for administrators to see them.

Real-time E-mail Collection

An agent sets up collection-policies on real-time and collects e-mails and then a manager can check the result directly on real-time. E-mails meeting the conditions can be collected through parameters such as e-mail sender, e-mail receiver, sending/receiving server, keywords of mail contents, attached files, etc. A manager selects attributes of files and sends queries to EFAB to have a real-time search for files satisfied with some specific conditions using EFGB. An e-mail's attributes such as keywords being included in contents of e-mails, IP address of sending this e-mail, mail server, receiver's e-mail address, assigned identification number, sending time, mailing program used, MIME format and code configuration, or encoding types are used in this case. After receiving queries, EFAB analyses a mail box (DBX) on real-time to search target e-mails (EML files) and then sends them to EFMB. Finally, EFMB sends these results to EFGB for administrators to see them.

Files and E-mails Collection

This is a function of collecting files and e-mails already collected and stored at the server by agents in advance. In case of file collection, a manager selects types of files to collect using EFGB. Options of collecting a file are a file's extension, a signature, and file generation time. Colleting policies EFGB selected are stored at user policy table by way of EFMB, and then are sent to EFAB through polling. EFAB searches target files from user's PC according to selected policies and extracts characters included in files. Additionally, EFAB combines extracted characters and attributes of a file to send them to EFSB and stores them as a type of ISAM file. EFSB combines characters included in received file and attributes of that file and stores them in MS-SQL database. In case of e-mail collection, a manager selects to index a mail box (DBX) stored at EFAB using EFGB. Colleting policies EFGB selected are stored at user policy table by way of EFMB, and then are sent to EFAB through polling. EFAB searches a mail box according to selected policies and extracts e-mails included in files and transforms as a type of EML. Additionally, EFAB sends EML-type messages to EFSB and EFSB parses received messages and stores them in MS-SQL database.

User's Abnormal Behavior Detection

This is a function of checking status of generating or deleting enterprise data. It chooses file extension, checking period, and number of times being checked through policy set-up, and checks how many target files were generated or deleted in this period. If generated or deleted files are very more than before, it decides this as an abnormal behaviour. A manager can set-up some policies for detecting abnormal behaviours using EFGB such as the starting time of detection or the standard number of generating or deleting a file, etc. EFMB checks monitoring logs stored at MS-SQL database in accordance with a managing function of enterprise data history through setting-up policies. The managing function of enterprise data history is required prior to starting this function.

Evidence Container Generation

This function is useful when a manager needs to store and manage specific files or e-mails separately. A manager generates DEB (Digital Evidence Bag) for storing required data after searching, stores and manages them independently. The second or the third departmentalized container can be generated using a web manager. A manager can set-up some policies for evidence containers among document files collected at the server using EFGB such as investigator for evidence container, kind/usage of evidence container, file list, the name of EFAB, etc. The evidence container can be made at EFSB through policies. ".tag" file stores some information of evidence container and ".index" file stores information of document files included at evidence container. These two files are always encrypted and compressed. ".bag" file plays a role of evidence container and contains the followings: investigator, kind/usage of container, the number of files, index type, the generation time of container, a file name, hash value of file, a file size, modification time, recent accessing time, etc.

Remote Forensic Image Generation and Transmission

This function generates an image of target agent's hard disk remotely to make the same copy of agent computer's hard disk for forensic analysis and sends it to the server. A manager can set-up some policies for generating forensic images remotely using EFGB such as a disk name or a scope of disk sectors. According to setting-up policies, EFAB generates an image of hard disks of user PC. The generated forensic image is not transmitted to EFSB in one image whole, but in a fixed size several times and EFSB stores that image. Finally EFSB stores attributes of transmitted forensic image at MS-SQL database such as a name of forensic image file, agent name of forensic image, creation time for forensic image, disk type, block size of image, starting and finishing sector of forensic image, etc.

Litigation Hold

This is a function of preventing users from modifying or deleting files or e-mails stored at agent computers from a specific time. It is surely needed to protect and preserve evidence data. In case of litigation hold for files, a manager can set-up files' attributes to prevent users from deleting these files which are satisfied with some conditions. Attributes data is transmitted from EFMB to EFAB. File attributes can be used are keywords included in contents of file and file metadata such as creating/modifying/recently accessing time for documents, file size, file name and extension, owner of file (account name), hash value of file, etc. In case of litigation hold for e-mails, a manager lets EFAB know whether it prevents users from deleting specific files or not.

File History Browsing

This selects target files having formats such as DOC, PPT, or PDF and sees history of target files' generation, modification, deletion, or duplication status. Specially, this function is very useful for forensic analysis since we can see attempting history of files having Litigation-Hold. A manager selects some extensions for monitoring agents' files using EFGB. Indexing policies, EFGB selected, are stored at EFSB's user policy table by way of EFMB and EFMB asks EFAB to get new policies using EFAB's polling. EFAB stores operation logs, such as creation, modification or deletion, of files having target extensions at MS-SQL database.

CONCLUSION

In this paper, we described component technologies for e-Discovery step by step and suggested agent-based information management and control system can manage ESI stored at some computers centrally and respond rapidly on demand, extracting discovery-related data using digital forensic technologies, as a prototype of preparing e-Discovery. This system could be used as a prototype of tools of coping with e-Discovery. Furthermore, if functions being mentioned above and some additional functions such as privacy information management including online data, high-level file control, relation analyses between data, contents-based search, meaningful log analysis, are added, it could be a powerful e-Discovery supporting solution.

ACKNOWLEDGMENTS

This work was supported by the IT R&D Program of MKE/KEIT[10035157, Development of Digital Forensic Technologies for Real-Time Analysis].

REFERENCES

Cohen, A., & Kalbaugh, G. (2010). ESI Handbook: Sources, Technology and Process. *Aspen Publishers*.

Container Format. (2011). Container format (digital). http://en.wikipedia.org/wiki/Container_format_(digital)

DLP. (2011). Data Loss Prevention. http://en.wikipedia.org/wiki/Data_loss_prevention_ software

ECA. (2011). Early Case Assessment. http://en.wikipedia.org/wiki/Early_case_assessment

EDRM. (2011). Electronic Discovery Reference Model. http://edrm.net

FRCP. (2006). Federal Rules of Civil Procedure. http://www.law.cornell.edu/rules/frcp

FRCP. (2011). Federal Rules of Civil Procedure. http://en.wikipedia.org/wiki/FRCP

Gartner. (2008). Market-Scope for E-Discovery Software Product Vendors".

Joachims, T. (1997). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. *Proceedings of the 14th International Conference on Machine Learning.*

Kim, Y., Hong, D., & Shin, S. (2010). FRCP and e-Discovery. *Weekly Technology Trends*.

Kim, Y., Shin, S., & Hong, D. (2010). Management, Identification and Preservation of ESI in e-Discovery. *Digital Forensic Technology Workshop*.

The Sedona Conference. (2011). http://www.thesedonaconference.org

Volonino, L., & Redpath, I. (2010). E-Discovery for Dummies. *Wiley*.

Yang, Y. & Peterson, J. (1997). A Comparative Study on Feature Selection in Text Categorization. *Proceedings of the 14th International Conference on Machine Learning.*