Avances en la Construcción de una Ontología para el Análisis Forense de Correo Electrónico

Progress in Building an Ontology for Forensics Email

Beatriz P. de Gallo
I.Es.I.Ing. /Facultad de Ingeniería
Universidad Católica de Salta
Salta, Argentina
bgallo@ucasal.edu.ar

Marcela Vegetti, Horacio Leone
INGAR/ Facultad Regional Santa Fe
Universidad Tecnológica Nacional
Santa Fé, Argentina
{mvegetti, hleone}@santafe-conicet.gov.ar

Resumen — ***La gran cantidad de información técnica resultante del análisis forense de un correo electrónico debe insertarse en el conjunto de pruebas documentales de la causa judicial que lo aborda. Esta documentación técnica debe ser accesible e interpretable por los profesionales de la criminalística y el derecho, y contar con un marco de referencia basado en la conceptualización formal del universo de discusión facilitaría esta accesibilidad. En particular, las ontologías resultan una herramienta de uso pluridisciplinar para facilitar el análisis de la prueba documental, por parte de todos los actores (abogados, jueces, investigadores y peritos) acuerden sobre el significado semántico de la documentación digital. Este trabajo tiene como objetivo describir los avances logrados en el desarrollo de una ontología que colabore en la interpretación de los datos producto del análisis forense de un correo electrónico.***

Palabras Clave – ontología; forensia digital; correo electrónico

***Abstract* — Forensic analysis of an email message produces a large amount of information that should be incorporated into the body of documentary evidence collected for the judicial action to which the message belongs. This technical documentation must be accessible to and interpretable by law and criminalistics professionals. A framework based on the conceptualization of the universe of discourse may facilitate such access. In particular, Ontologies play an important role as multidisciplinary tools which help different actors (lawyers, judges, investigators and experts) reach an agreement on the semantic meaning of the information in the documentary evidence. This paper describes progress in the development of an ontology to assist in the interpretation of technical data obtained from the forensic analysis of an email.**

Keywords - ontology; digital forensics; email

I. INTRODUCCIÓN

Cada día toma mayor importancia el uso del correo electrónico, en su carácter de ***registro formal de una comunicación entre dos partes***. En el contexto legal, durante los últimos años ha crecido exponencialmente la presentación de correos electrónicos como prueba documental en las causas judiciales[1].

Si bien la Forensia Digital avanzó en concordancia con la tecnología, es necesario aún trabajar un aspecto que no es propiamente del ámbito tecnológico y que genera un conjunto de interrogantes que impactan grandemente en los resultados que se obtienen, i.e., la ***interpretación de los resultados***.

El volumen de datos que se obtiene al realizar el análisis forense debe ser interpretado a la luz de la pesquisa. La cantidad de información técnica resultante del análisis de un correo electrónico debe insertarse en el conjunto de pruebas documentales de la causa judicial, colocándolo en un estadio de lectura que facilite la interpretación de esos datos técnicos por parte de los profesionales de la criminalística y el derecho. Se requiere mucho más que la identificación de una dirección IP (Internet Protocol) o la trazabilidad del correo electrónico. Hoy en día se exige que estos datos se presenten ***sistemáticamente*** y ***semánticamente*** en el marco de la causa judicial, no como información técnica, sino como dato documental.

En el contexto de este requerimiento "no técnico", se encuentra la motivación de este trabajo. Resulta conveniente contar con un marco de referencia basado en la conceptualización formal del universo de discusión. Y en

1 De cada 10 causas judiciales en los que se solicitó la participación de un perito informático, 26 % tratan acerca de la autenticación y autoría de correos electrónicos. (Fuentes propias).

particular, las ontologías resultan una herramienta universal o pluridisciplinar para facilitar el análisis de la prueba documental, por parte de todo los actores (abogados, jueces, investigadores y peritos).

Este trabajo tiene como objetivo describir el avance logrado en el desarrollo de una ontología que colabore en la interpretación de los datos producto del análisis forense de un correo electrónico.

La organización de este trabajo es la siguiente: la sección 2 describe el marco teórico sobre el objeto de estudio (correo electrónico) señalando los aspectos de interés para el análisis forense y refiere brevemente el marco teórico de las ontologías. En la sección 3 se muestran los avances logrados a la fecha en la construcción de una Ontología para el Análisis Forense de Correos Electrónicos. Finalmente, en la sección 4 se presentan las actividades en curso y a desarrollar para concluir el trabajo.

II. TRABAJOS RELACIONADOS

A. *El Correo Electrónico*

1) Definición y características propias

La literatura específica contiene diferentes definiciones del término correo electrónico. A los fines del presente trabajo, puede tomarse como válida la señalada por las Directrices de la Unión Europea 2002/58/CE [1] relativas a la protección de datos, en las que se definen el Correo Electrónico como "*Todo mensaje de texto, voz, sonido o imagen enviado a través de una red de comunicación pública que puede almacenarse en la red o en el equipo terminal del receptor hasta que éste acceda al mismo*".

El correo electrónico (o e-mail) se ha transformado en el medio de comunicación más utilizado en el tráfico de red facilitando grandemente la comunicación entre las personas. Además de acortar tiempos y distancia, permite el intercambio de múltiples tipos de datos (video, imagen, audio) y se encuentra accesible en prácticamente todos los medios de comunicación tecnológicos, habiendo avanzado rápidamente en la telefonía celular. Esta última característica de "portabilidad" abre instancias de comunicación que antes no estaban presentes, reforzando la inmediatez de la comunicación interpersonal, con el agregado de que ahora existe ***un registro de esta comunicación***. Si bien en el correo electrónico se adopta lenguaje coloquial, y es muy utilizado para la comunicación informal, es importante reconocer que es posible recuperar la conversación y utilizarla como prueba de que tal comunicación existió.

Desde el punto de vista legal, el correo electrónico tiene interés como documento probatorio en un juicio, por lo que resulta importante introducirlo con la fuerza y el rigor técnico suficiente para que actúe en el proceso judicial de igual manera que lo hiciera cualquier otra prueba material.

Es importante señalar que la característica de volatilidad de los datos digitales, impacta negativamente en el reconocimiento de un correo electrónico (o cualquier otro componente digital) como prueba documental. Los profesionales del foro judicial quieren "ver" la prueba, darle forma, buscar el origen, su historia, imaginar todo lo que hay alrededor de este componente, cual si fuera un "arma homicida". Recién en esta instancia pueden reconocer la validez de la prueba digital, i.e., una vez que logran ver la "sustancia" y "consistencia" de elemento probatorio.

2) Naturaleza Jurídica del correo electrónico

Relacionado con el contexto jurídico, Castro Bonilla [2] menciona tres enfoques o miradas diferentes que se pueden considerar en un correo electrónico:

a) Como correspondencia o comunicación: con idéntica naturaleza que el correo postal tradicional, se encuentra protegido por las leyes que regulan la correspondencia epistolar. Tanto los datos recibidos cuanto los enviados desde la cuenta de correo, constituyen elementos protegidos bajo el principio de inviolabilidad de las comunicaciones. En nuestro país, la Ley 26.388 incluye el término "correo electrónico" en el tipo de violación de correspondencia privada establecida por los arts. 153 y 154 del Código Penal.

b) Como conjunto de datos: al ser datos personales, su manipulación se encuentra supeditada a las normas relativas a la protección de datos personales. A partir del análisis forense de un correo electrónico es posible identificar una serie de datos del individuo (receptor/emisor del correo) que vulnera el derecho a la autodeterminación informativa de una persona. En Argentina, la Ley 25.326 "…*tiene por objeto la protección integral de los datos personales asentados en archivos, registros, bancos de datos, u otros medios técnicos de tratamiento de datos, sean éstos públicos, o privados destinados a dar informes, para garantizar el derecho al honor y a la intimidad de las personas, así como también el acceso a la información que sobre las mismas se registre,*…".

c) Como transmisor de material protegido por derechos de autor: al ser un medio de transmisión de datos de gran diversidad de formatos, hace posible la difusión de material protegido por derechos de autor, con un impacto notorio en la vulnerabilidad de la propiedad intelectual del material contenido o adjunto en un correo electrónico.

Debe considerarse también los tipos de correos electrónicos desde el punto de vista de la propiedad de la cuenta y las implicancias de esta identificación en el contexto legal. Reyes [3] identifica dos tipos de propietarios de la cuenta de correo electrónico:

a) Privado: es el medio por el cual un usuario en forma directa posee una cuenta proporcionada por algún proveedor de servicios (yahoo, hot mail, etc.), sea ésta en forma gratuita o mediante un pago, pero en ambos casos queda supeditado a las norma de seguridad y de uso de la cuenta que acepta en el momento de realizar la suscripción. Este tipo de correo es estrictamente personal y, por ende, está protegido como una correspondencia inviolable, no pudiendo en principio establecerse ningún tipo de excepción, lo que significa que no puede ser manipulado, interceptado, intervenido o alterado de alguna forma si no se posee una autorización por parte del

receptor o por medio de una orden judicial, pues está protegida su intimidad como un derecho fundamental.

b) Empresarial y/o Laboral: es el medio por el cual la empresa ha contratado un dominio propio o cuenta, y en cuya dirección figura su nombre o las iniciales, que habitualmente se identifica con el dominio de la cuenta, de tal forma que queda plenamente identificada la empresa en cada correo enviado por el empleado.

No hay un consenso generalizado respecto de si en este último caso es legal inspeccionar una cuenta de correo corporativo. Por una parte, el fallo del Tribunal Europeo de Derechos Humanos [4], justifica la inspección por una empresa de los mensajes privados de los empleados si son emitidos desde cuentas corporativas.

En nuestro país, en un reciente caso [5], se establece la nulidad de la prueba de correos electrónicos corporativos aportada por una empresa, por considerar que vulneraba la privacidad del empleado, toda vez que fue realizado sin la debida autorización de juez competente.

Estas acciones de nulidad son ocasionadas mayormente por no respetar el marco jurídico correspondiente para la *obtención* de la prueba, cayendo en la teoría de "los frutos del árbol envenenado" en clara referencia a que las pruebas de un delito obtenidas de manera ilícita están viciadas y son prueba nula. Así, un tema de especial interés en las pericias informáticas es el procedimiento o protocolo para la obtención, preservación y presentación de la prueba. En nuestro país, existen acordadas de la justicia provincial de Neuquén [6] y Rio Negro [7] que implementa protocolos de actuación para el tratamiento de pruebas digitales. Cuando no se cuenta con un marco legalmente aprobado para el tratamiento de la prueba digital, los peritos forenses informáticos acuden a los procedimientos fijados para la recolección de la prueba física definido en la justicia local más la aplicación de las buenas prácticas profesionales avaladas por las principales asociaciones internacionales abocadas a temas de seguridad informática.

3) Estructura y componentes

A los fines del presente trabajo, se abordará brevemente aquellos aspectos del correo electrónico que resulten de interés para un análisis forense.

Tomando como base la tipificación propuesta por Banday [8] para el análisis forense de un correo electrónico se puede identificar tres componentes principales: los actores participantes en la transmisión, la arquitectura lógica y la estructura interna de un correo electrónico.

En referencia a los ***actores participantes*** se puede decir que, si bien la comunicación de un correo electrónico requiere de personas que actúan como emisor y receptor del mensaje, no son éstos los únicos partícipes de la transmisión. Banday establece un mapa de relaciones y caminos posibles que puede recorrer un correo durante el proceso de transmisión e identifica los procesos responsables de sostener el servicio –denominados *actores*- que actúan internamente durante la transmisión.

La ***arquitectura lógica*** de funcionamiento del envío/recepción de mail consta de varios componentes de hardware y software. Además de los dispositivos utilizados por el *Autor/Receptor*, sean éstos una computadora, un celular o cualquier otro equipo utilizado para el acceso al sistema de correo electrónico, se requiere de objetos transparentes al usuario principal que actúan en el proceso de creación, envío, transmisión, entrega y lectura del e-mail, como por ejemplo: componentes de la capa inferior de TCP/IP (routers, modem) y de la capa de aplicación de TCP/IP (nodos de e-mail) así como los diversos protocolos y paquetes de software utilizados (SMTP, HTTP, clientes de correo, etc.).

En cuanto a la ***estructura interna,*** un correo electrónico contiene un conjunto de *campos* asociados en *layers* que cumplen una función específica en la transmisión de datos. De ellos interesa identificar en particular: la cabecera del mensaje (con los metadatos requeridos para la identificación del correo), el cuerpo o contenido, los archivos adjuntos asociados al mensaje, y las direcciones IP involucradas en la transmisión.

Sobre este tema en particular, se trabajó en un caso ejemplo [9] para estudiar la arquitectura señalada por Banday.

4) Análisis Forense de un Correo Electrónico

Una pericia [10] es un conjunto de operaciones técnicas científicas puestas en práctica para el esclarecimiento de un posible hecho ilícito y ordenadas por el Tribunal interviniente.

En cuanto a los ***puntos de pericia***, su ofrecimiento permitirá al Juez determinar la procedencia de la prueba, es decir, la congruencia entre los aspectos a conocer y la necesidad de un técnico para que lo asesore. Los puntos periciales se proponen en un pliego que señala las cuestiones técnicas, de manera clara y precisa, siempre referidas al tema que se dilucida en la litis y que técnicamente puedan ser respondidas por el Perito[2].

Usualmente los puntos de pericia referidos a correos electrónicos abordan las siguientes cuestiones:

- Verificar la autenticidad, dirección, fecha y hora de emisión y recepción, autoría si fuera posible de los correos electrónicos cuyos impresos se adjuntan, así como cualquier otra información que estime relevante.
- Determinar la validez de origen de los e-mail enviados por A
- Expedirse acerca de la existencia de los envíos de los correos electrónicos emitidos entre A y B
- Informar si las cuentas de correo electrónico a@dominio.com y b@dominio.com están o estuvieron activas en los períodos consignados
- Informar si A tiene como correo electrónico la cuenta a@dominio.com
- Informar si en fecha dd/mm/aaaa desde la cuenta a@dominio.com se envió un correo a la cuenta b@dominio.com y en ese caso transcriba el texto del mensaje.

[2] A los fines del presente trabajo, se tomará en consideración el contexto más usual de la Forensia Digital, es decir, el ámbito de la justicia, sin perjuicio de que lo dicho aquí se pueda aplicar en los restantes ámbitos en que actúa esta disciplina (investigaciones privadas, fraudes tecnológicos internos, etc.).

- Llevar a cabo la exploración y análisis de todos los soportes informáticos idóneos para el almacenamiento de información (notebooks, netbooks, pc's, discos rígidos, diskettes, cd's, dvd's, pen drives), instalados o que se encuentren en determinado lugar, ello con el objeto de determinar el soporte informático desde donde se hayan confeccionado documentos y/o remitido e-mails que dieran origen a las actuaciones.
- Realizar una correlación de eventos que relacione los datos de los correos electrónicos recibidos (dirección IP, emisor, destinatario/s, asunto, fecha, hora y observaciones varias) con los datos informados por la empresa X (nombre del usuario, titular del dominio, documento del titular, domicilio y localidad).
- Determinar si los correos que aparecen como enviados o recibidos son de las fechas que indica la parte en su demanda y si son coincidentes con los que mencionan como prueba la parte actora.
- Determinar si es posible que los mensajes enviados o recibidos, como los textos adjuntos que allí figuran, pueden haber sido alterados en sus fechas y horas de emisión o recepción.
- Identificar cuáles fueron los equipos de origen y de destino del mensaje, si los servidores o proveedores de Internet – que operan a modo de enlace entre el remitente y el destinatario del e-mail- cuentan con información inherente a la fecha en que se produjo el envío o reenvío del e-mail, duración de conexión, archivos adjuntos, número de identificación de los equipos de origen y destino de las comunicaciones – datos de tráfico realizado.

Con independencia de la particularidad de cada punto de pericia, toda la información vinculada a un correo electrónico que se incorpora como prueba, se sostiene si es posible probar dos cosas: la autenticidad y la existencia del correo electrónico.

Los elementos que permiten verificar la ***autenticidad*** de un correo electrónico son los siguientes:

- la identificación de los datos del remitente (nombre de usuario, cuenta de correo y dirección IP),
- la trazabilidad del mismo (diferentes servicios o agentes que intervienen en la transmisión), y
- los datos del destinatario (nombre de usuario, cuenta de correo y dirección IP.

En cuanto a la ***existencia*** de un correo electrónico, ésta se puede probar fehacientemente cuando se comprueba la presencia del archivo digital del mismo tanto en el dispositivo emisor (o en el servidor del ISP[3] del emisor) como en el dispositivo receptor del correo (o en el servidor ISP del receptor); y ambos archivos digitales son idénticos.

Desde el punto de vista de la forensia digital, existen muchas técnicas y herramientas que ayudan al Perito Informático en el análisis de un correo electrónico.

En cuanto a las *técnicas*, la investigación forense de un correo electrónico se puede abordar desde varias ópticas: el análisis de los datos de cabecera, el análisis de los equipos emisores/receptores del correo, el análisis de los servidores ISP, el análisis de los metadatos ocultos en las aplicaciones utilizadas para la escritura del correo. Todas estas técnicas se utilizan habitualmente en conjunto para la confirmación redundante sobre la autenticidad del correo electrónico.

Existen muchas y diversas herramientas disponibles para analizar un correo electrónico que fueron analizadas en [11], a los citados allí se agrega el trabajo de Devendran, Shahriar y Clincy [12] acerca de un estudio comparativo de varios software open source para el análisis de correos electrónicos.

La elección de la técnica y herramientas más adecuadas se deduce de la estrategia de investigación que siga el perito, la cual dependerá de ciertos factores: dispositivo a analizar (PC, celular, servidor, etc.); cliente de correo (residente en el dispositivo o web mail); cantidad de correos (se debe analizar toda la cuenta o solo un correo determinado) y facilidad de acceso a la prueba (acceso al email enviado y al recibido, solo a uno de ellos, al servidor de correo, etc.)

Si bien las técnicas y herramientas mencionadas constituyen el marco formal y científico que califican la profesionalidad y rigor metodológico que se requiere en un análisis forense, los resultados que se obtienen no siempre cumplen su cometido: brindar información fundada sobre los puntos en litigio, o mejor dicho, responder los puntos de pericia de manera clara y contundente.

El principal inconveniente radica en las dificultades que tienen los partícipes no informáticos de la causa (jueces, fiscales, abogados, investigadores forenses de otras disciplinas) para *interpretar los datos técnicos* a la luz de la causa judicial, y en el contexto del resto de las pruebas documentales presentes en el litigio.

De allí que se requiera de un sistema de representación que haga posible mostrar los datos en función del objetivo que se persigue (expresado en los puntos de pericia) y vinculados semánticamente en base a la relación que los mismos mantienen entre sí.

B. Las Ontologías

1) Definiciones básicas y características

Existe abundante bibliografía a la que se puede recurrir para tomar los conceptos iniciales sobre ontologías. La definición de Gruber [13] parte del concepto de conceptualización como una abstracción o visión simplificada del universo que queremos representar con algún propósito. Las ontologías son una representación explícita o formal de esa conceptualización, recurriendo a la representación de los diferentes elementos que conforman el universo de discusión (objetos, conceptos y otras entidades), así como las relaciones que los vinculan.

Las ontologías proponen un marco de referencia basado en el conocimiento, mediante un vocabulario de representación que describe cada elemento según una definición declarativa y axiomas formales que acotan la interpretación y permiten una aplicación correcta de esos términos.

En el conjunto de sistemas de representación del conocimiento, las ontologías se definen según sus características distintivas:

[3] ISP = Internet Service Provider, proveedor del servicio de internet.

- Permiten consensuar el significado de los elementos y relaciones de un universo de discusión, de manera que es posible desarrollar un software para modelar los procesos de toma de decisiones por parte de los gestores del conocimiento.
- Abordan siempre un dominio acotado del conocimiento. Si bien uno de los principales problemas al definir una ontología es identificar *donde está el límite* de lo que queremos representar, esa misma acotación sustenta y valida la representatividad de la ontología. Es decir, una vez definido el dominio, las reglas de representación de una ontología permiten modelar acabadamente ese ámbito restringido denominado *universo de discusión*.
- Se recurre a la lógica formal para representar los componentes mediante los conceptos tradicionales de *objetos, clases, instancias, restricciones y propiedades*. Al utilizar modelos formales para la representación, es posible la aplicación de lenguajes compatibles con entornos abiertos y comprensibles para una máquina, tales como OWL, RDF, XML.

Se recurre a la *semántica* como hilo conductor para definir los componentes que se representarán, así como las relaciones de vinculación entre ellos, permitiendo expresar el dominio en base al significado que tienen sus componentes en el marco de referencia en el que actúan.

2) Metodologías y herramientas para la construcción de ontologías

En la Ontology Summit 2007 [14] se discutió sobre la gran variedad de metodologías de diseño, algunas de las cuales enfatizan una que otra propiedad en la construcción de la ontología: la ingeniería de requerimientos o la evaluación y validación de la aplicación informática resultante. Incluso se proponen metodologías sin diseño como en el caso de las folsonomías que parten del comportamiento local de miles de individuos.

No debe perderse de vista el concepto en sí de una metodología: como herramienta solo es útil en la medida en que su uso acompaña el logro del objetivo propuesto. Y en el caso particular de las ontologías semánticas, en las que la definición del dominio es una parte sustancial, se debe orientar la construcción de la ontología según sea la característica distintiva del tema.

Así, las ontologías que tratan sobre vocabularios o taxonomías deben reforzar las instancias de significación de las palabras en el dominio que están abarcando; en otros casos, como en ontologías de integración de datos, es de interés profundizar la etapa de validación de los metamodelos de datos; o, en contextos específicos como la Forensia Digital, cobra vital importancia la validación de las instancias de captura de la prueba digital y su correspondiente "cadena de custodia".

En particular, interesa el trabajo de Corcho et al. [15] en el que presentan una adaptación al dominio legal español de una taxonomía de clases sobre entidades legales, aplicando la metodología METHONTOLOGY y la herramienta WebODE.

Esta metodología propone guías de actividades para la especificación, conceptualización, formalización, implementación y mantenimiento de la ontología a construir, bajo un esquema de procesos iterativos que ayudan en el ajuste del modelo a construir. A continuación se sintetizan estas fases:

- La actividad de ***especificación*** permite determinar por qué se construye la ontología, cuál será su uso, y quiénes serán sus usuarios finales.
- La actividad de ***conceptualización*** se encarga de organizar y convertir una percepción informal del dominio en una especificación semi-formal, para lo cual utiliza un conjunto de representaciones intermedias (RRII), basadas en notaciones tabulares y gráficas, que pueden ser fácilmente comprendidas por los expertos de dominio y los desarrolladores de ontologías.
- La actividad de ***formalización*** se encarga de la transformación de dicho modelo conceptual en un modelo formal o semicomputable.
- La actividad de ***implementación*** construye modelos computables en un lenguaje de ontologías (Ontolingua, RDF Schema, OWL, etc.).
- La actividad de ***mantenimiento*** se encarga de la actualización y/o corrección de la ontología, en caso necesario.

METHONTOLOGY también identifica actividades de gestión (planificación, control y aseguramiento de la calidad), y de soporte (adquisición de conocimientos, integración, evaluación, documentación y gestión de la configuración).

Existen varios trabajos de investigación que relacionan ambos temas: las ontologías y el análisis forense. Para el caso particular de los correos electrónicos, es de interés el trabajo de Balakumar et al. [16] sobre la definición de una ontología para la clasificación y categorización de e-mail con el objetivo de conformar un filtro para la detección de spam mediante la conformación de una whitelists de remitentes conocidos, así como el trabajo de clasificación de e-mails propuesto por Taghva [17] en referencia a la exigencia legal de resguardar ciertos registros de datos residentes o anexados a correos electrónicos. De los últimos avances sobre el tema se puede considerar de interés el trabajo de Alzaabi et al. [18] que propone *F-DOS* un conjunto de ontologías que modelan formalmente el contenido de un teléfono inteligente.

III. ONTOLOGÍA PARA EL ANÁLISIS FORENSE DE CORREO ELECTRÓNICO

En esta sección se presenta una ontología que permite definir la semántica de los conceptos relacionados con el análisis forense de un correo electrónico a fin de formalizar los mismos, así como sus relaciones y restricciones. Siguiendo la guía de actividades de Methontology, se mostrarán los avances logrados.

1) Fase de Especificación

La primera actividad a desarrollar consiste en determinar el objetivo de la ontología, su funcionalidad y destinatarios finales. El objetivo de una ontología para el análisis forense de correos electrónicos es el de construir un marco de referencia formal y científico, que sustente el análisis semántico de un

correo electrónico en su carácter de prueba documental, basando esa interpretación en la relación que esos datos tienen en el contexto de la causa.

La función de esta herramienta, será la de servir de apoyo a los profesionales forenses no informáticos, facilitando la interpretación de los datos obtenidos con foco en la relación que los mismos tienen con el contexto de análisis y otras pruebas documentales, mediante la identificación de los conceptos, atributos, valores y relaciones que mantienen estos datos, más allá de las características técnico-informáticas que se obtienen como resultado del análisis forense de los correos electrónicos.

Los destinatarios finales son aquellos usuarios no informáticos, o con escasa experiencia con la tecnología (abogados, jueces, fiscales, otros investigadores forenses), que necesitan interpretar los resultados del análisis forense de un correo electrónico en el marco de la causa judicial, y a quienes –usualmente- los datos técnicos "crudos" que se obtienen como resultado de un análisis forense digital les resultan complejos o difíciles de comprender.

La definición del dominio parte de un conjunto de preguntas que guían la delimitación o acotación del universo de discusión, y que ayudan a decidir qué objetos son relevantes y cuales no son representativos para el análisis.

Estas preguntas tienen como cometido los siguientes:

- establecer las funcionalidades a las que debe responder la aplicación informática resultante de la ontología;
- delimitar el dominio tal como lo conciben los usuarios finales; y
- considerar los puntos de vista de los usuarios a la hora de modelizar los conceptos de la realidad.

Los interrogantes base de esta ontología se pueden buscar en los puntos de pericia que usualmente se proponen al solicitar un análisis forense de un correo electrónico y que se han enunciado en el apartado anterior, de allí se extractan las ***preguntas de competencia*** en lenguaje natural:

1. ¿Cuáles son las partes de un correo electrónico que resultan de interés para un análisis forense?
2. ¿Cuáles son los componentes informáticos a través de los cuales se escribe y se lee un correo electrónico?
3. ¿Cuáles son los datos o componentes que permiten validar la existencia de un correo electrónico? Esta pregunta puede descomponerse en las siguientes:
   3.1. ¿Cuál es la fecha, hora y dirección IP de emisión del correo electrónico?
   3.2. ¿Cuál es la fecha, hora y dirección IP de recepción del correo electrónico?
4. Dado un correo electrónico ¿Cuáles son los datos que permiten identificar la autoría y recepción del mismo? Esta pregunta puede descomponerse en las siguientes:
   4.1. ¿Cuál es el nombre de usuario y dirección de e-mail del Autor del mismo?
   4.2. ¿Cuál es el nombre de usuario y dirección de e-mail del Receptor del mismo?
   4.3. ¿Es posible establecer el seguimiento del mensaje desde que se envía hasta que se recibe?
   4.4. ¿Cuáles son los diferentes actores/servicios que participaron de la transmisión?

En una nueva iteración del proceso de construcción de la ontología, se está trabajando ahora en la recolección de información mediante una encuesta dirigida a los peritos informáticos, con el fin de obtener más instancias de los puntos de pericia, y *validar* las preguntas de competencia definidas.

2) Fase de Conceptualización

En esta fase, se trabaja sobre la definición del dominio, a partir de la percepción informal inicial que se tiene del mismo, estructurándolo en un modelo conceptual que luego sea posible formalizarlo, recurriendo a herramientas de representaciones intermedias (tablas y gráficos).

Se definió el *vocabulario de términos* utilizado como base para responder a las preguntas de competencia.

Tomando los componentes descriptos en el apartado anterior, los términos de interés deben surgir de los tres conjuntos definidos: los actores participantes en la transmisión, la arquitectura lógica y la estructura interna de un correo electrónico.

Siempre teniendo presente que las preguntas de competencia apuntan al carácter probatorio del correo electrónico (en cuanto a su incorporación como prueba en un expediente judicial), en función de las características de autentificación y existencia citadas anteriormente, se definieron los conceptos, atributos de instancia y relaciones más notables.

La Tabla 1 describe los conceptos y atributos de instancias identificados.

TABLE I. CONCEPTOS Y ATRIBUTOS DE INSTANCIA

Concepto	Descripción	Atributos de instancia
Correo Electrónico	Todo mensaje de texto, voz, sonido o imagen enviado a través de una red de comunicación pública que puede almacenarse en la red o en el equipo terminal del receptor hasta que éste acceda al mismo	Cuenta de correo emisor Cuenta de correo receptor ID del Mensaje Fecha y hora emisión Fecha y hora recepción Dirección IP Asunto Cuerpo Archivo adjunto
Emisor	Persona que envía el correo electrónico	Alias Nombre Firma
Receptor	Persona que recibe el correo electrónico	Alias Nombre Firma
Hardware	Componente informático que interviene en el emisión/transmisión/recepción del correo electrónico	Tipo de Hardware Identificación única del equipo Dirección IP del equipo Nombre del equipo
Equipo emisor/ receptor	Hardware utilizado por el usuario para la escritura/lectura del correo electrónico	Identificación única del equipo Descripción del equipo Dirección IP del equipo Ubicación geográfica del equipo
Dispositivo de Transmisión	Hardware que interviene en el proceso de transmisión del correo electrónico	Identificación única del dispositivo Dirección IP del dispositivo Nombre del dispositivo Descripción del dispositivo Ubicación geográfica del dispositivo

Concepto	Descripción	Atributos de instancia
Servidor de correo	Servidor que almacena el correo electrónico enviado/recibido	Identificación única del servidor Dirección IP del servidor Dominio del servidor Descripción del servidor Ubicación geográfica del servidor
Gestor de servicio	Componente de hardware o software que participa en la emisión/transmisión/recepción del correo	Tipo de servicio
Software	Aplicación informática que interviene en la emisión/transmisión/recepción del correo electrónico	Identificación única del software Nombre del software Descripción del software
Cliente de correo	Aplicación informática que interviene en la escritura-emisión y lectura-recepción del correo electrónico	Identificación única del cliente de correo Nombre del cliente de correo Descripción del cliente de correo
Expediente	Contenedor de documentos sobre la causa judicial	Id_expediente Carátula_expediente Juzgado

3) Fase de Formalización

Esta etapa implica que el modelo conceptual especificado debe representarse por medio de un lenguaje formal, dando lugar a los componentes de la ontología: clases, atributos, conceptos, relaciones, funciones, axiomas e instancias. Una vez identificados los términos, se seleccionaron aquellos que actúan como *conceptos*, para armar la ***Taxonomía de Conceptos***, que se muestra en la Figura 1.

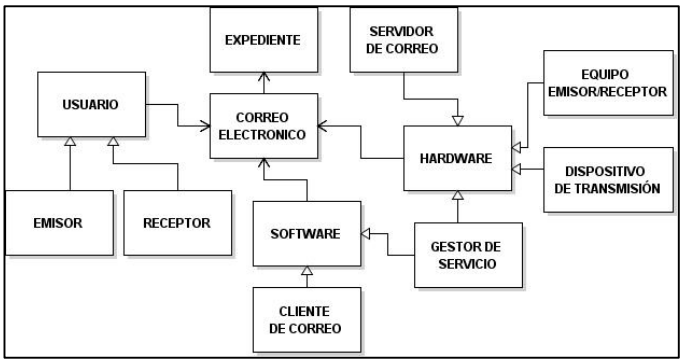

Figure 1. Principales Conceptos

La clase "software" representa los diversos programas que intervienen en la emisión / transmisión / recepción de un correo electrónico, de ellos resulta de interés la subclase "clientes de correo" ya que permiten vincular el correo electrónico con una persona (usuario emisor/receptor).

En cuanto a la especialización de la clase "hardware", en particular se identifica como sub-clase a los equipos utilizados para la emisión/recepción del correo, los servidores de correo y los dispositivos que participan de la transmisión del correo electrónico. Por su parte los equipos de emisión/recepción colaboran en la identificación de la vinculación usuario-correo electrónico, en los servidores de correo se encuentran almacenados los correos electrónicos, mientras que los dispositivos de transmisión muestran la *trazabilidad* del mensaje.

Partiendo de la narrativa escrita en lenguaje natural –los puntos de pericia- se establecen las ***relaciones ad hoc*** existentes entre los conceptos definidos en la taxonomía. Las relaciones deben establecer con exactitud y precisión, indicando el origen y destino de cada una, evitando imprecisiones o sobreespecificación de esos puntos. La Figura 2 muestra el proceso principal en el análisis forense, señalando los conceptos que participan y la vinculación que existe entre ellos.

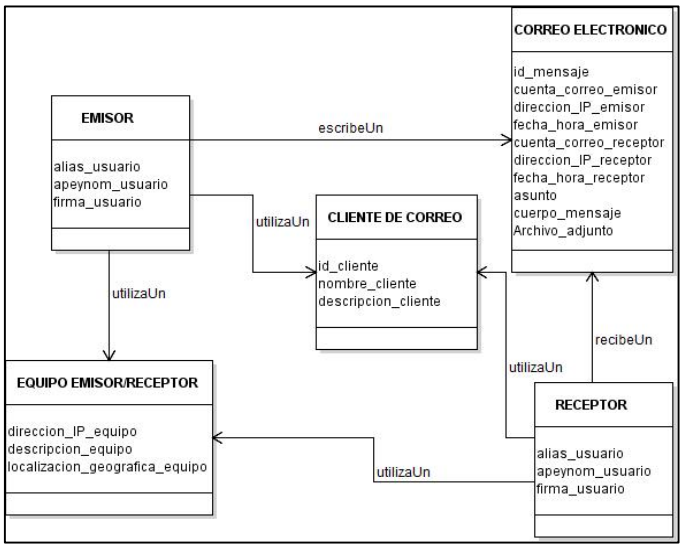

Figure 2. Relaciones ad-hoc para emisión/recepción del correo electrónico

Una pregunta reiterada a los peritos es acerca de cómo es posible aseverar con plena certeza que el correo escrito por una persona es el que efectivamente recibió la otra (y viceversa), esto se puede responder si se establece el procedimiento de comunicación y la vinculación entre todos los componentes y actores que participan de la transmisión. La relación ad-hoc que permite observar la vinculación de los componentes intermedios durante el proceso de transmisión se muestra en la Figura 3.

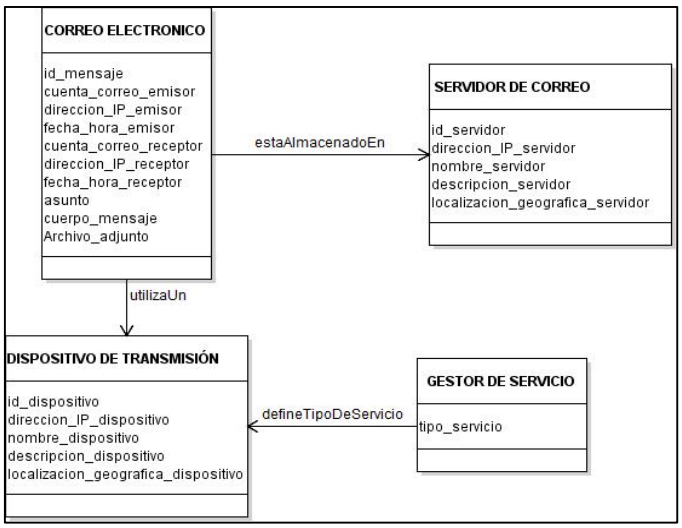

Figure 3. Relaciones ad-hoc para para transmisión del correo electrónico

Mientras que la Figura 4 muestra la vinculación entre los distintos partícipes de un correo electrónico, cuando se realiza un envío masivo, a través de una lista de distribución de correos.

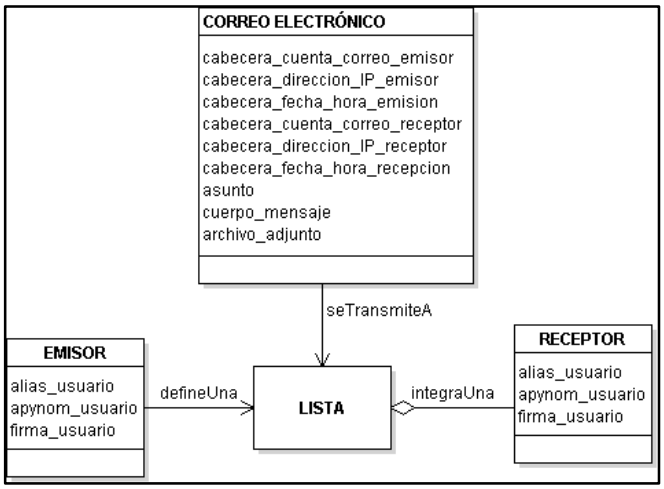

Figure 4. Relaciones ad-hoc para para la Lista de Distribución de correo electrónico

Luego de identificadas las relaciones ad-hoc, METHONTOLOGY propone como siguiente paso la generación de una serie de tablas que definen el conjunto de metadatos de la ontología: el Diccionario de Conceptos, la Tabla que describe las Relaciones Binarias ad-hoc, los atributos de instancia y de clase, y las constantes. Por razones de espacio no se incluyen en el presente trabajo pero pueden verse en [11].

A partir de las preguntas de competencia, se avanzó en la definición de los axiomas, y que en cierta forma, conforman la base del análisis forense de un correo electrónico:

- ***Axioma 1: sobre la autenticidad de un correo electrónico***
  Un correo electrónico es auténtico cuando se identifican: los datos del remitente (nombre de usuario, cuenta de correo y dirección IP), la trazabilidad del mismo (diferentes servicios o agentes que intervienen en la transmisión) y los datos del destinatario (nombre de usuario, cuenta de correo y dirección IP).
- ***Axioma 2: sobre la existencia de un correo electrónico***
  Un correo electrónico existe cuando se comprueba la presencia del archivo digital del mismo tanto en el dispositivo emisor (ó en el servidor del ISP del emisor) como en el dispositivo receptor del correo (ó en el servidor ISP del receptor); y ambos archivos digitales son idénticos.

4) Fase de Implementación

Para esta fase se eligió *Protègè*[4], un editor de ontologías de código abierto. La Figura 5 muestra la jerarquía de clases y las relaciones de la versión de prueba que actualmente se está trabajando.

Al momento de iniciar la etapa de validación del modelo construido, se buscó un conjunto de instancias ejemplos que permitieran desarrollar las primeras corridas de consistencias, advirtiendo entonces que la estructura de almacenamiento de los metadatos que utilizan los gestores de correo no resultan fácilmente accesibles. Estos archivos se generan en los más variados formatos, desde .mbx que es un archivo de texto plano fácilmente legible con un editor de notas, hasta estructuras complejas como .pst que son legibles solo desde "dentro" del propio cliente de correo (como es el caso de Microsoft Outlook). En [19] se detalla la investigación realizada al respecto, concluyendo en que es necesario realizar una depuración semántica inicial del corpus de correos obtenido a partir de la extracción con herramientas forenses.

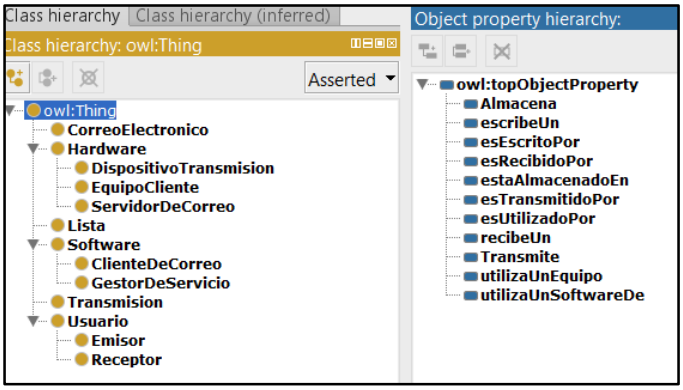

Figure 5. Jerarquía de Clases

El siguiente avance en el proyecto fue la búsqueda de herramientas para el preprocesamiento de los datos, como preparación previa para poblar la ontología. En ese sentido, se investigó acerca de los aportes que la Minería de Datos podría hacer a este proyecto.

A lo largo del tiempo, la minería de datos se ha dotado de diferentes herramientas que ayudan al tratamiento de los datos –no ya desde el enfoque transaccional- sino más bien desde la necesidad de generar un contexto de trabajo analítico.

La DM se ha provisto de importantes herramientas no sólo para el análisis sino también para la búsqueda y recuperación de información de interés, que se encuentra almacenada en corpus documentales de datos no estructurados

Wen-Yang Lin [20] sintetiza en un esquema muy claro la relación de DM y las ontologías durante el proceso de gestión del conocimiento, vinculando ambos temas a partir de la multidimensionalidad y de la especialización de los reservorios de datos. El autor destaca el rol de las ontologías como marco de referencia para la representación de los datos en todas las etapas del proceso de KDD[5].

En cuanto a la aplicación de la minería de datos en relación a los correos electrónicos, existen varios antecedentes que se pueden tomar como referencia[6].

La construcción de una ontología contempla la validación del modelo instanciado desde un contexto real y acorde a los objetivos que se persiguen con la misma. Cuando las instancias

[4] Ver 5.0.0 (Build beta-23)

[5] KDD: (Knowledge Discovery from Databases) Descubrimiento de Conocimiento en Bases de Datos

[6] En el trabajo [19] se detalla exhaustivamente la investigación realizada

provienen de repositorio con datos mayormente no estructurados (como los correos electrónicos), es importante recurrir a los procedimientos y herraumientas de búsqueda y recuperación de información o *Information Search and Retrieval (ISR)* [7]. En base a la problemática cada vez más compleja para poblar las ontologías con información cierta y pertinente, surge un nuevo campo de estudio que se aboca a la *Instanciación Automática de Ontologías* (IAO) a partir de textos o datos no estructurados.

Según sea el tipo de datos que almacena el reservorio que representa el dominio (datos estructurados o no estructurados), se presentará en mayor o menor grado la necesidad de utilizar técnicas automáticas o semiautomáticas de extracción de datos para obtener un conjunto que luego permita la instanciación de la ontología.

Son varios los autores que mencionan la necesidad de recurrir a tecnologías de recuperación de la información para poblar las ontologías [21], [22], [23], [24] marcando la necesidad de desarrollar herramientas inteligentes y métodos de extracción de datos para el conocimiento (procesamiento de metadatos, identificación y categorización de conceptos claves, detección primaria de inconsistencias, técnicas de búsquedas automáticas o semiautomáticas, etc.).

El proceso de adquisición del conocimiento para la población de una ontología comprende tres etapas: recuperación de la información, extracción de la información necesaria y carga de la ontología. La primera etapa consiste en encontrar los documentos que contengan la información requerida para la ontología. Aquí es de mucha utilidad la categorización de textos, que identifica subconjuntos de textos relacionados por un criterio predefinido, aunque también puede recurrirse a otras herramientas de minería de textos como las redes bayesianas, las redes neuronales o las máquinas de vectores soporte. La etapa de extracción de la información necesaria se ocupa de encontrar los objetos, clases, instancias, restricciones y propiedades que conforman el corpus de la ontología.

Desde esta perspectiva, es esencial realizar adecuadamente la selección, limpieza, enriquecimiento, reducción y transformación de datos, como paso inicial de validación de la información entrante. Así, se ha trabajado identificando:

- *datos incompletos (*atributos faltantes o incompletos, datos fuera de rango):
- *datos con ruido* (fallas en las rutinas de validación de los datos de ingreso, limitaciones tecnológicas)
- *datos inconsistentes* (ajuste de códigos o conversiones de tipo al provenir de diferentes fuentes, tuplas duplicadas).

La minería de datos propone diversas técnicas para esta etapa, como ser, la utilización de valores estadísticos para resolver la ausencia de valores en un atributo, clustering para valores anómalos o fuera de rango o la agregación y normalización para la transformación de los datos.

[7] ISR: encontrar material (generalmente documentos) de naturaleza no estructurada (generalmente de texto) que satisface una necesidad de información desde dentro de grandes colecciones (generalmente almacenada en las computadoras).

La Figura 6 muestra el diagrama de procesos semiautomáticos para la extracción y limpieza de datos de una casilla de correos electrónicos, realizado a partir de la utilización de la herramienta *RapidMiner*[8], tomando como ejemplo una vista parcial de la ontología que se está trabajando.

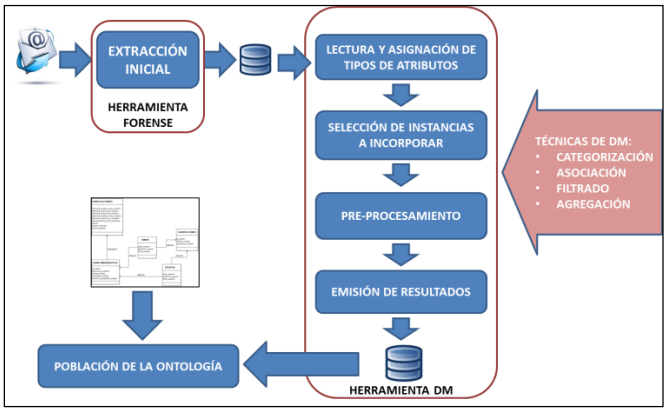

Figure 6. Diagrama de Procesos Semiautomáticos para la extracciòn y limpieza de datos

En particular, se utilizarán los conceptos y asociaciones involucrados en la relación ad-hoc para la emisión/recepción del correo electrónico que se muestran en la Fig. 2. En [19] se describe como se aplicó este procedimiento para un caso ejemplo de una cuenta de correo con 1162 mensajes.

La primera actividad consistió en realizar la extracción de todos los correos electrónicos obrantes en la casilla de interés. Esta ***Extracción Inicial*** permitió la obtención de todos los datos necesarios para la instanciar la ontología. Una vez obtenido el corpus de correos electrónicos, se inició el trabajo de procesamiento con herramientas propias de DM.

Por empezar, de manera automática es posible asignar el tipo de dato y otras características necesarias (identificación del campo ID, identificación de campo etiqueta para procesos de predicción, etc.) en el mismo proceso de lectura del archivo. Esta etapa es la denominada ***Lectura y asignación de tipos de atributos*** en el proceso señalado en la Figura 6.

Para la ***Selección de Instancias a Incorporar*** se recurrió a las técnicas de DM que sean necesarias para conformar el conjunto de instancias que se utilizarán posteriormente para poblar la ontología. En esta etapa es cuando se agrega valor al proceso de instanciación, ya que –usualmente- esta identificación de información relevante se realiza *manualmente* y con la presencia del experto.

En la etapa de ***Pre-procesamiento*** se recurrió a los componentes para depuración y limpieza de datos que habitualmente están incorporadas en las herramientas de DM, y que en el caso de los procesos de actualización de las ontologías se trabajan desde las técnicas de ISR. Según sea el caso, se puede trabajar con técnicas de Categorización, Asociación, Filtrado, Agregación, entre otras. En el caso ejemplo que se cita se utilizaron técnicas de asociación, filtrado y agregación.

[8] https://rapidminer.com/

Por último, en la etapa de ***Emisión de Resultados*** se obtuvo el conjunto de datos seleccionados y depurados que poblarán la ontología, en formato de archivo tabular.

A la fecha se está trabajando en la construcción de una aplicación que permita visualizar las instancias y poblar la ontología. Para la construcción de la aplicación se está investigando acerca de métodos WebML [25] para la especificación y el análisis de la navegación en sistemas software.

IV. CONCLUSIONES

El plan de trabajo prevé una serie de acciones iterativas tendientes a validar la consistencia de la propuesta en varios aspectos:

- Luego de concretar la implementación de la ontología será necesario realizar nuevos ciclos de formalización e implementación, validando el modelo a partir de un mayor conjunto de instancias y casos.
- Se espera difundir la encuesta sobre puntos de pericia de correos electrónicos, con la suficiente cobertura como para obtener casos que sean suficientes para ajustar las preguntas de competencia.
- Debe ajustarse las pruebas experimentales sobre el proceso de extracción y limpieza que incluyan las variantes necesarias para probar el modelo, como ser: la comprobación con más casos periciales como el citado, e incluso de diferente contenido al del ejemplo; la cobertura plena del modelo de la ontología, i.e., a todas las clases, propiedades y relaciones definidas en la misma; y la inclusión de otras técnicas de DM que también pueden ser de utilidad para la depuración previa de los datos (la categorización de textos, por ejemplo).
- Resulta conveniente avanzar en la búsqueda e integración de ontologías existentes así como en el desarrollo del contexto jurídico (representado aquí con el concepto de *expediente)* en el cual actuará finalmente la ontología propuesta.

AGRADECIMIENTOS

Este trabajo ha sido financiado en forma conjunta por CONICET, la UTN (PID 25-O156) y el Consejo de Investigaciones de la Universidad Católica de Salta.

REFERENCIAS BIBLIOGRÁFICAS

[1] DIRECTIVA 2002/58/CE DEL PARLAMENTO EUROPEO Y DEL CONSEJO SOBRE LA PRIVACIDAD Y LAS COMUNICACIONES ELECTRÓNICAS, 12 de julio de 2002, p.37

[2] Castro Bonilla, A."El uso legítimo del correo electrónico", *II Congreso Mundial de Derecho Informático*, 2002

[3] Reyes Reyes, Luz Adina. Ponderación de lo regulado en el artículo 196 bis del Código Penal Costarricense y la intervención del patrono en la revisión del correo electrónico del trabajador, *Universidad Estatal a Distancia de Costa Rica*, Tesis de Maestría, http://repositorio.uned.ac.cr/reuned/handle/ 120809/1436, (2015).

[4] CASE OF BĂRBULESCU v. ROMANIA. *European Court of Human Rights (Fourth Section)*, http://hudoc.echr.coe.int/ eng?i=001-159906, STRASBOURG, 12 January 2016

[5] G., R. S. y otros, : Cámara Nacional de Apelaciones en lo Criminal y Correccional, sala I, 13/02/2015, AR/JUR/20660/2015/

[6] Acuerdo N° 4908, Protocolo de Actuación para Pericias Informáticas, Poder Judicial de la Provincia de Neuquén, http://boficial.neuquen.gov.ar/pdf/bo12090703315.pdf, 2012

[7] Acordada 05/2014 del SUPERIOR TRIBUNAL DE JUSTICIA DE LA PROVINCIA DE RIO NEGRO, http://tsjrn.opac.com.ar/pgmedia/Acordadas/2014-005_AC.pdf, 2014

[8] Banday, M. Tarik, "TECHNIQUES AND TOOLS FOR FORENSIC INVESTIGATION OF E-MAIL*", International Journal of Network Security & Its Applications (IJNSA),* Vol.3, No.6, November 2011

[9] Rivetti E., Parra H.B., "Verificación de la trazabilidad de un correo electrónico mediante un caso ejemplo", Cuadernos de Ingeniería 2015,. Número 9 del 2015. ISSN 2422-6572 (On line), ISSN 2422-6564, in press.

[10] Fernández, Eduardo Enrique: "Aspectos legales del peritaje". *Revista INDICIOS*, Año 2. Vol. 2. La Rioja (Argentina) 2011. pp. 24-33.

[11] Gallo Beatriz P. de, Vegetti Marcela, Leone Horacio, "Ontología para el Análisis Forense de Correo Electrónico", CoNaIISI 2014 Actas del 2° Congreso Nacional de Ingeniería Informática/Sistemas de Información, San Luis, Argentina, ISSN: 2346-9927, 2014

[12] Devendran, Vamshee Krishna, Hossain Shahriar, and Victor Clincy. "A Comparative Study of Email Forensic Tools." *Journal of Information Security*6.2 (2015): 111.

[13] Gruber, Thomas R., "A Translation Approach to Portable Ontology Specifications", *Appeared in Knowledge Acquisition*, 5(2):199-220, 1993.

[14] Ontology Summit 2007 Communiqué, 2007, version 1.0.0 / 2007.04.24

[15] Corcho, Óscar y Fernández-López, M. y Gómez-Pérez, A. y López-Cima, A., "Construcción de ontologías legales con la metodología METHONTOLOGY y la herramienta WebODE", *Law and the Semantic Web. Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications. Springer-Verlag*, pp. 142-157. ISBN 0302-9743, 2005

[16] Balakumar, M., & Vaidehi, V. (2008, January). Ontology based classification and categorization of email. In *Signal Processing, Communications and Networking, 2008. ICSCN'08. International Conference on* (pp. 199-202). IEEE

[17] Taghva, K., Borsack, J., Coombs, J., Condit, A., Lumos, S., & Nartker, T. (2003, April). Ontology-based classification of email. In *Information Technology: Coding and Computing, International Conference on* (pp. 194-194). IEEE Computer Society.

[18] Alzaabi, Mohammed. "The Use of Ontologies in Forensic Analysis of Smartphone Content." *Journal of Digital Forensics, Security and Law* 10.4 (2015): 105-114.

[19] Gallo Beatriz P. de, Vegetti Marcela, Leone Horacio, "Población de ontologías con datos no estructurados utilizando herramientas de minería de datos", CoNaIISI 2015 Actas del 3° Congreso Nacional de Ingeniería Informática/Sistemas de Información, Buenos Aires, Argentina, ISBN: 978-987-1896-47-9, 2015

[20] Wen-Yang Lin, Ontology-Based Data Mining A Case in Multidimensional Association, Dept. of Computer Science and Information Engineering , National University of Kaohsiung, 2006

[21] Fierros, J. D. G. TESIS DE MAESTRÍA EN CIENCIAS, Poblado Automático de Ontologías Espaciales a Partir de Texto no Estructurado, *Centro Nacional de Investigación y Desarrollo Tecnológico, Departamento de Ciencias Computacionales, Cuernava, Mexico, 2012*.

[22] Paredes Moreno, A. (2007). Técnicas de depuración e integración de ontologías en el ámbito empresarial.

[23] Cala A., Schorlemmer M, Noriega P., PROTOTIPO DE UN MODULO DE BUSQUEDA SEMANTICA PARA LA PLATAFORMA GreenIDI. TR--IIIA--2013--01, *IIIA-CSIC Barcelona, 2013*

[24] Daly, M., Grow, F., Peterson, M., Rhodes, J., & Nagel, R. L. (2015, April). Development of an automated ontology generator for analyzing customer concerns. In *Systems and Information Engineering Design Symposium (SIEDS), 2015* (pp. 85-90). IEEE.

[25] Cuaresma, María José Escalona. *Modelos y técnicas para la especificación y el análisis de la navegación en sistemas software*. Diss. Universidad de Sevilla, 2004.