электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

 # 10, октябрь 2015

УДК 004.457

Исследование файлов текстовых форматов на примере форматов RTF, DOC, DOCX

***Шавловский А.Б.**, студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Юриспруденция, интеллектуальная собственность и судебная экспертиза»

Научный руководитель: Безик О.В., ассистент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Юриспруденция, интеллектуальная собственность и судебная экспертиза»
o.bezick@bmstu.ru

Информационные технологии являются быстро развивающейся сферой жизни современного общества. Почти каждый день в мире появляются новые виды компьютеров, электронных устройств, программного обеспечения и др. Недаром многие ученые считают наш современный социум постиндустриальным или так называемым «информационным обществом».

Однако бурный рост информационных технологий не только способствует улучшению жизни людей, но и подвергает опасности, если идет речь об использовании этих технологий в неправомерной деятельности.

Постоянно растет количество преступлений, совершаемых при помощи компьютеров и иных электронных устройств. Перед судебными экспертами ставятся все новые и новые задачи, решение которых требует от них глубоких познаний в структурах файлов и принципе действия тех или иных устройств и технологий. В данной статье рассмотрен вопрос, связанный с исследованием файлов текстовых форматов на примере форматов RTF, DOC, DOCX.

Общие задачи исследования файлов текстовых форматов

Пользователи персональных компьютеров, электронных устройств, создают и обмениваются файлами текстовых форматов, созданными посредством таких приложений, как Microsoft Office, Apache OpenOffice, Kingsoft Office и другие. Из таких файлов можно извлекать различные данные, такие как имя автора документа, дата и время

последних изменений, размер документа. В то же время файлы могут содержать и очень важные сведения, к примеру, данные о ранее удаленном или измененном тексте.

Так, к примеру, файлы текстовых форматов, созданные в приложениях Microsoft Office, по умолчанию хранят в себе следующую информацию, которая в зависимости от задач, стоящих перед экспертом, может оказаться весьма значимой [2]:

- Дата создания файла;
- Дата создания содержимого;
- Автор (условное имя)
- Количество редакций
- Дата последнего сохранения
- Дата последнего вывода на печать
- Общее время редактирования и прочее.

Исходя из следственно-судебной практики, в последнее время увеличивается количество изъятой правоохранительными органами различной компьютерной техники, электронных носителей информации. При обнаружении на этих изъятых устройствах различных файлов текстовых форматов, назначается судебная компьютерно-техническая экспертиза.

При исследовании компьютерно-техническим экспертом файлов текстового формата, возникает необходимость извлечения всех данных о создании, изменении либо удалении файла (в том числе текста в самом файле), включая установление:

- Даты и времени начала и конца работы над документом;
- Общего времени редактирования документа;
- Структура документа (самостоятельных частей) и способа ее формирования;
- Даты печати документа;
- Выявление на электронных носителях так называемых «документов-родителей» и «документов-потомков».

Наличие дополнительной технической информации и ее количество зависит от вида файла текстового формата и его структуры.

Файлы текстовых форматов, как правило, бывают двух видов:

- Не структурированные – это файлы, содержащие в себе только, текст, а иногда еще и отдельные управляющие символы. Пример: файлы с расширением «.txt»;
- Структурированные – это файлы, содержащие в себе пользовательскую (текст, размер, разметка), а также дополнительную, служебную информацию, так называемые «метаданные».

Формат файла – это правило размещения в теле файла пользовательской и служебной информации, так называемых метаданных. Формат файла иногда указывается в его имени, как часть, отделённая точкой. Обычно эту часть называют расширением имени файла. Метаданные – это информация о данных. Если эта информация оформлена отдельно, то она будет считаться данными как самостоятельный объект, метаданными в отношении некоторых данных [9].

Файл, имеющий формат, как правило, имеет заголовок, последовательность байт (2, 4 и более) в начале тела файла, которая сигнализирует о файле некоторого конкретного формата. В криминалистической литературе заголовок называют сигнатурой файла, что не вполне корректно.

Если описание формата файла опубликовано и доступно каждому – это открытый формат файла. Если же описание формата файла запатентовано разработчиком и не находится легально в открытом доступе, то это проприетарный (закрытый) формат файла.

Один из видов проприетарных форматов файла является формат Rich Text Format, RTF — межплатформенный формат хранения текстовых документов, предложенный как мета-тэговый формат для редактора Word в 1982 году группами программистов, основавших компании Microsoft и Adobe [5]. В целом же, Rich Text Format – это универсальный формат, который обязан отображаться при открытии его любой программой просмотра текстов в любой операционной системе. С тех пор спецификация формата несколько раз изменялась. После разрыва отношений с Microsoft компания Adobe продолжила самостоятельное развитие метатэгового языка, заложенного в основу RTF, создав в 1985 году язык PostScript.

Файлы формата RTF имеют следующий синтаксис:

<File> '{' <header> <document> '}'

Файл состоит преимущественно из команд управления, воспринимаемых программой чтения RTF-файлов, которые могут быть управляющими словами и операторами. Перед командой вводится обратная косая черта « \ ».

RTF файл состоит из групп команд управления, обрамленных фигурными скобками. В группах могут содержаться или описываться разные типы данных: текстовые блоки, графические объекты, таблицы, исполняемые файлы, служебная информация и др.

Некоторые операторы, которые представляют криминалистическую значимость в вопросах установления обстоятельств создания файла:

- \info – начало блока информации, в котором могут быть следующие операторы:
- \title – заголовок

- \author – имя автора (условное, обычно – имя учетной записи в ОС семейства Windows)
- \operator – имя последнего исполнителя (условное, по аналогии с именем автора файла)
- \company – компания
- \keywords – ключевые слова
- \versionN – версия текста (N – номер редакции)
- \vernN – внутренняя версия файла (англ. Internal Version Number)
- \creatim – дата и время создания документа;
- \revtim – дата и время последней корректировки документа;
- \printtim – дата и время последней распечатки документа;
- \buptim – дата и время последнего сохранения документа;
- \edmins – общее время редактирования в минутах
- \revision – отметка об изменениях;
- \revtbl – таблица с информациях о редакциях текста
- \insrsid – отметка о вставленном тексте;

Наиболее популярным форматом файла, наряду с RTF, является двоичный формат файлов Microsoft Office Word 97-2007(*.doc), основанный на структуре OLE 2.0

OLE — технология связывания и внедрения объектов в другие документы и объекты, разработанная корпорацией Microsoft. OLE-файл состоит из так называемых виртуальных потоков. Рассматриваемые файлы имеют собственную «файловую систему внутри» схожую по свойствам с FAT или NTFS [4].

Виртуальный поток – это данные, которые читаются как линейный поток, хотя их физическое расположение в файле может быть фрагментировано. Это могут быть данные пользователя или метаданные.

OLE-файл построен как файловая система. Пространство файла разбито на сектора, размер которых определен при создании файла и, как правило, равен 512 байтам. На рис. 1 показаны виртуальные потоки файла «andrey.doc» при исследовании его утилитой olebrowse.

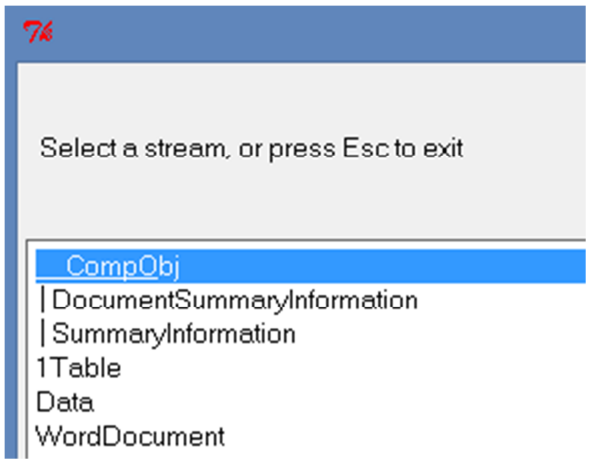

Рис. 1. Структура файлов формата «doc» на примере файла «andrey.doc» в утилите olebrowse

«WordDocument» – главный или основной поток, содержащий текст документа в кодировке Unicode с обратным порядком байтов (менее значимый байт хранится первым, более значимый последний) [6].

Потоки «\SummaryInformation» и «\DocumentSummaryInformation» в совокупности содержат метаданные, они воспринимаются как единый массив.

Поток «1 Table» – содержит служебную информацию, которая связана с изменениями содержимого документа пользователем (вставка рисунка, сокращение документа) или автоматически (автосохранение документа) [3].

Поток «\CompObj» содержит данные, необходимые для функционирования самого механизма OLE.

Начиная с Microsoft Word версии 2007 по умолчанию для новых документов используется формат «.docx».

Файл с расширением «.docx» - документ, созданный с помощью приложений Microsoft Word 2007 и более новой версии (2010, 2013). Формат файла основан на Open XML, который состоит из ZIP архива, содержащего XML и двоичные файлы. Данные из файла могут быть проанализированы без его модификации, путем распаковки архива и непосредственного исследования содержимого файла. На рис. 2 показано содержимое файла «andrey.docx» при распаковке архива.

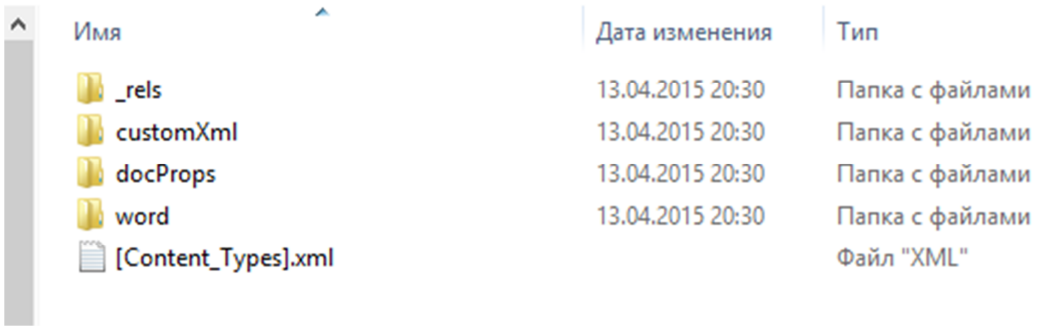

Рис. 2. Содержимое файла «andrey.docx» при распаковке архива

Файл _rels/.rels содержит информацию о структуре документа. Он содержит пути к метаданным так же, как и основной XML файл с содержимым файла.

В каталоге docProps содержатся два файла app и core, которые в совокупности содержат следующие метаданные:

app.xml:

- <TotalTime>mmm</TotalTime>, где mmm – общее время редактирования в минутах;
- <Application>MOW</Application>, где MOW – приложение, с помощью которого создавался файл;
- <Company>name</Company>, где name –условное название компании, на которую зарегистрировано программное обеспечение;
- <AppVersion>15.0000</AppVersion>, где 15.000 – версия приложения;

core.xml:

- <dc:creator>Andrey</dc:creator> и <cp:lastModifiedBy> Admin</cp:lastModifiedBy>, где Andrey и Admin – условные имена пользователя, создавшего и изменившего файл соответсвенно;
- <cp:revision>8</cp:revision> , где 8 – количество редакций файла;
- <dcterms:created xsi:type="dcterms:W3CDTF">xxx</dcterms:created> <dcterms:modified xsi:type="dcterms:W3CDTF">yyy</dcterms:modified>, где xxx и yyy – время создания и последнего изменения файла соответственно.

Практическое исследование файла текстового формата

Зачастую файлы текстовых форматов становятся важным источником значимой для рассмотрения дела информацией. С правовой точки зрения особенность анализа файлов текстовых форматов состоит в том, что такое исследование позволяет доказать причастность определенных лиц к совершению преступлений, в том числе и через

информационно-телекоммуникационную сеть Интернет. Поэтому такая информация может использоваться как в уголовном, так и в гражданском судопроизводстве.

Однако, одновременно с высоким уровнем развития технологий усложняется структура файлов, что важно учитывать для того, чтобы иметь возможность в полной мере и безошибочно анализировать данные.

Выбор инструментальных средств для исследования файлов текстовых форматов не так велико. На сегодняшний день в экспертной практике РФ методики исследования файлов текстовых и смежных форматов не существует.

Рассмотрим некоторые вопросы исследования таких файлов. В качестве примера был взят файл, содержащий текстовую информацию, а также некоторую, скрытую для глаза обычного пользователя, данные. Было проведено исследование текстового файла «test.doc». На первом шаге использовалась программа для извлечения метаданных MetaDataMiner Catalogue Pro (версия 4.2.34) разработчика Soft Experience, которая извлекает метаданные из файлов Microsoft Office [1]. На рис. 3 представлен скриншот этой программы и результаты исследования файла «test.doc».

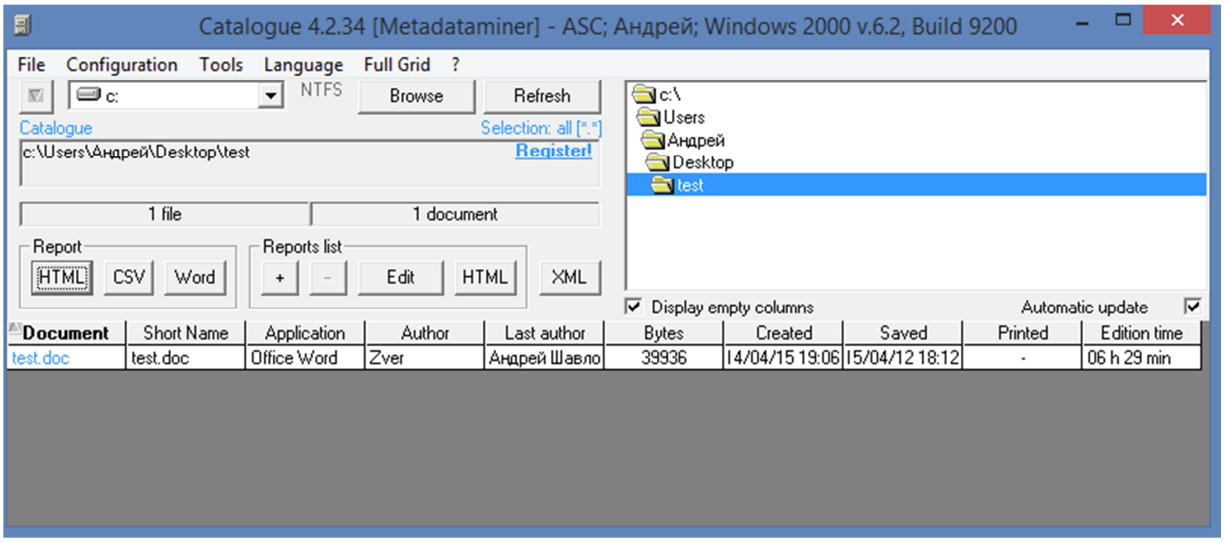

Рис. 3. Исследование файла «test.doc» программой Metadataminer Catalogue Pro 4.2.34

Основной особенностью этой программы является извлечение метаданных из всех файлов, находящихся в каталоге, который можно выбрать в правой части окна программы. Кроме того в настройках программы можно указать перечень данных, необходимых для извлечения и отображения. На рис. 4 показан скриншот этих настроек программы Metadataminer Catalogue Pro 4.2.34.

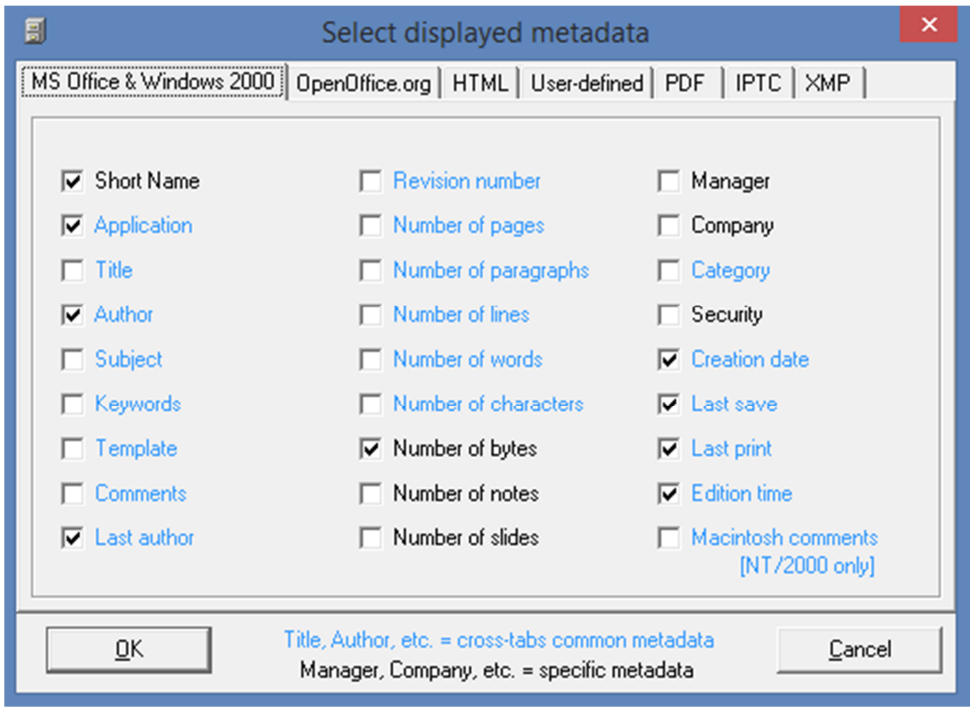

Рис. 4. Окно выбора метаданных в окне программы Metadataminer Catalogue Pro 4.2.34

Результат извлечения метаданных можно сформировать в отдельный отчет и поместить в файл doc, html или csv. На рис. 5 представлен скриншот с извлеченными метаданными из исследуемого файла в файл приложения Word.

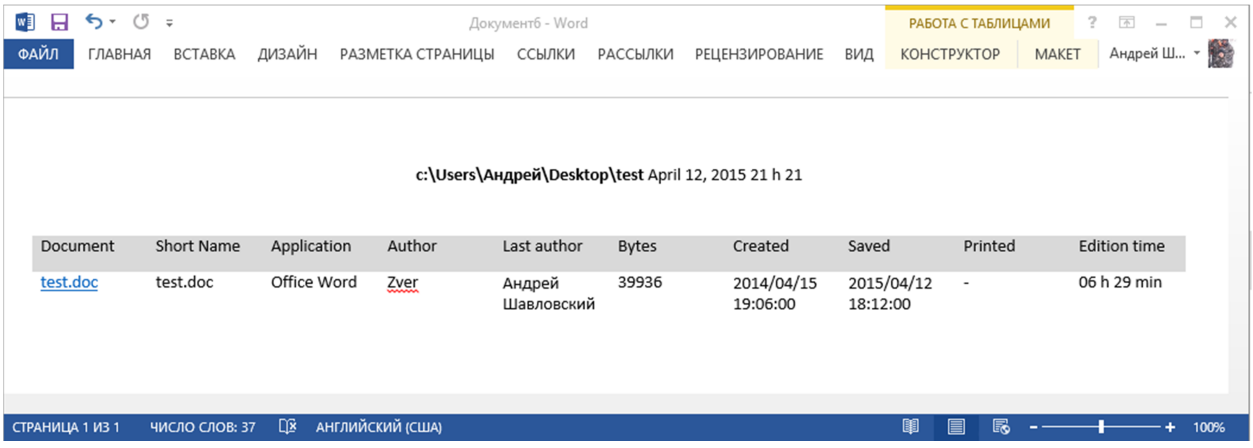

Рис. 5. Исследованные метаданные из файла «test.doc», экспортированные в файл Word

Следующим шагом было проведено исследование структуры файлов текстового формата при помощи программы Olebrowse, написанная на языке Python. Для её запуска требуется интерпретатор этого языка. Программа обладает понятным графическим интерфейсом. При запуске утилита просит указать путь к файлу формата OLE2, который следует исследовать. После этого программа анализирует его и показывает виртуальные

потоки файла. Каждый из потоков утилита может сохранить как отдельный файл, а также представить в шестнадцатиричном виде.

Виртуальные потоки можно сохранить в отдельные файлы, если в структуре файла обнаружен скрытый виртуальный поток, его можно проанализировать, чтобы определить формат файла, который в этом потоке спрятан [8]. Для эксперимента был исследован файл «test.doc». На рис. 6 показан скриншот утилиты olebrowse при исследовании файла «test.doc».

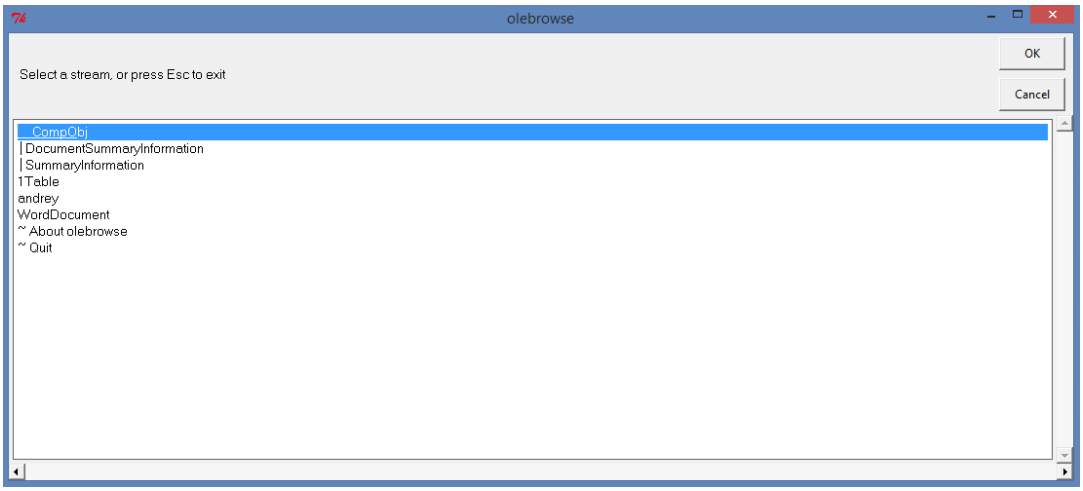

Рис. 6. Исследование файла «test.doc» программой olebrowse

Как видно, программа olebrowse вывела список потоков файла «test.doc», из которых «нестандартным» потоком является поток «andrey».

Далее поток «andrey» был сохранен в отдельный файл для его последующего исследования, это представлено на рис. 7.

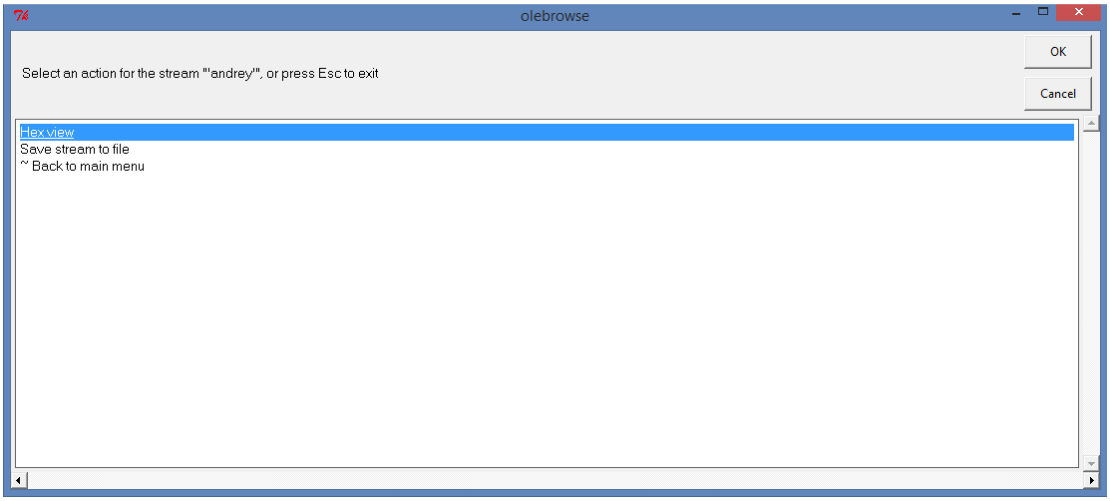

Рис. 7. Сохранение обнаруженного скрытого потока «andrey» в отдельный файл

Ранее сохраненный поток «andrey» из программы olebrowse в отдельный файл «andrey.bin», был исследован при помощи программы FileType Verificator, которая была разработана белорусской компанией Shedko Software. Данная программа предназначена для определения типов файлов по их содержимому. Она понятна и достаточно проста в использовании. На рис. 8 приведен скриншот этой программы при исследовании файла «andrey.bin».

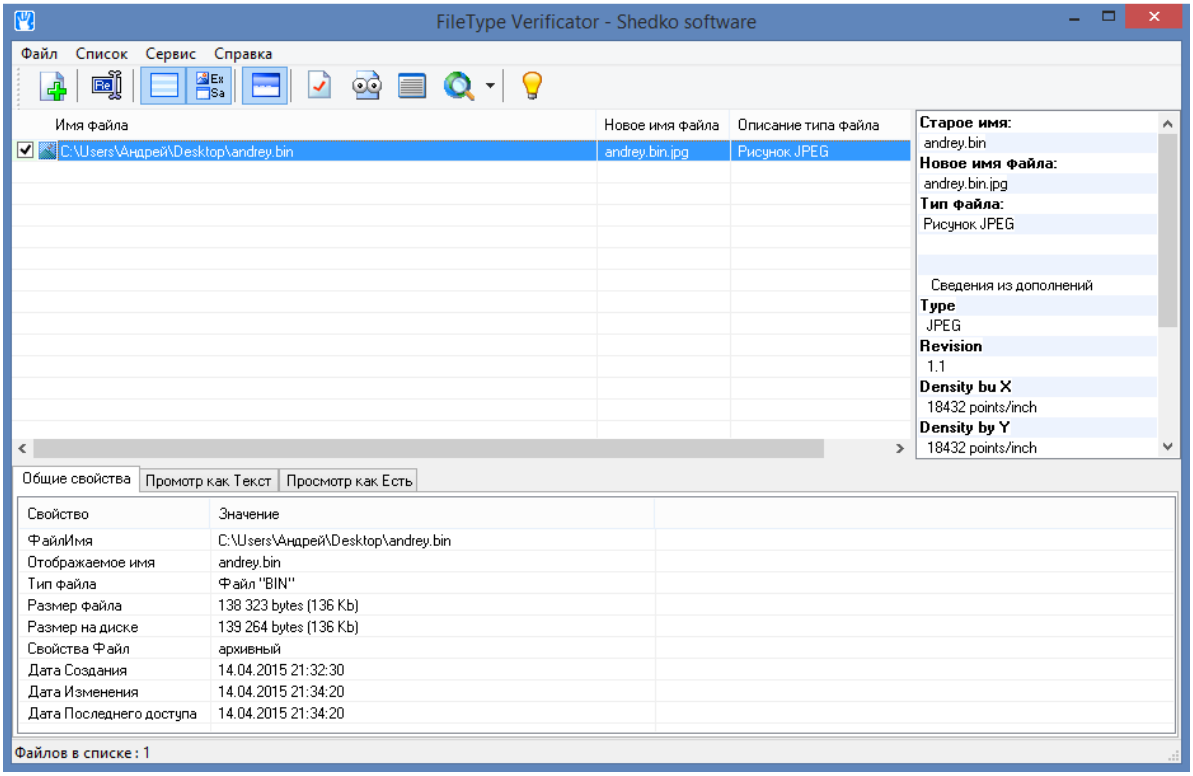

Рис. 8. Исследование файла «andrey.bin» программой FileType Verificator

Удалось получить данные о дате создания и изменения этого файла, его размере, а также, что в этом файле скрыто изображение формата JPEG, которое также при помощи программы FileType Verificator можно просмотреть и сохранить в отдельное изображение. На рис. 9 представлен скриншот с изображением, которое было обнаружено программой в файле.

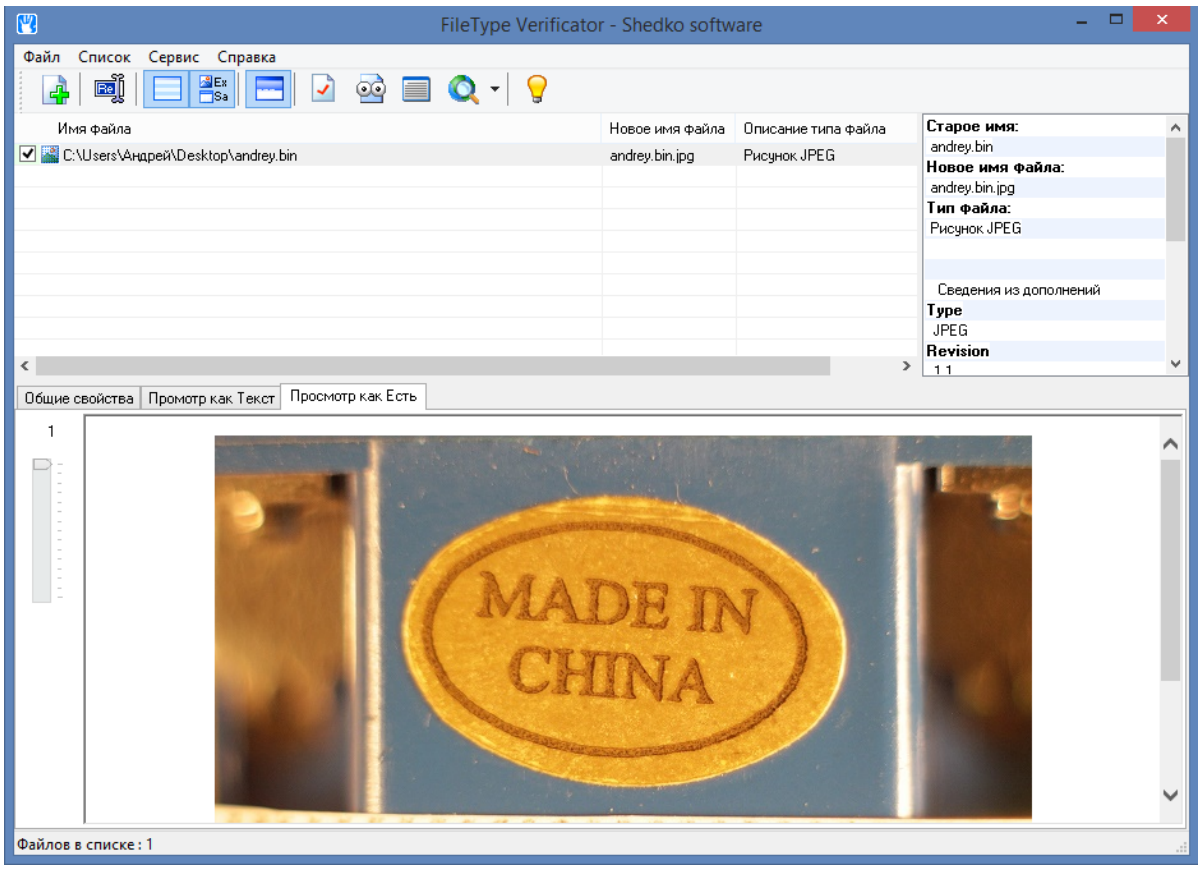

Рис. 9. Изображение, которое находилось в потоке «andrey.bin»

Таким образом, при исследовании файла «test.doc» мы не только получили метаданные из него, но и обнаружили в нем скрытый поток, который содержал файл формата JPEG.

Заключение

В рамках данной работы были рассмотрены структуры отдельных файлов текстовых форматов. Необходимость понимания этой структуры очень важна при работе с этими файлами, особенно когда они становятся непосредственными объектами судебных компьютерно-техническим экспертиз.

Была проанализирована методика исследования текстовых файлов с помощью программ и утилит, находящихся в открытом доступе. В качестве практического примера я провел исследование файла формата «.doc» (структура OLE2) на извлечение из него метаданных и обнаружение скрытых виртуальных потоков. Так был обнаружен такой скрытый виртуальный поток, который был дополнительно исследован и в нем было обнаружено изображение.

Данное исследование наглядно показало, как злоумышленники могут использовать текстовые файлы в своих интересах для совершения преступных действий, пересылки

запрещенной информации, сокрытия каких-либо сведений о преступной деятельности. Все это непосредственно связано с работой судебного компьютерно-технического эксперта, который в своей деятельности преследует цели осуществления правосудия, изобличения преступников, доказывания вины или ее отсутствия отдельных лиц.

Список литературы

1. Описание расширения «.doc», чем открыть файл doc. Режим доступа: www.open-file.ru/types/doc (дата обращения 15.03.2015).
2. Описание файла DOC, Википедия. Режим доступа: www.ru.wikipedia.org/wiki/DOC (дата обращения 18.03.2015).
3. Общие сведения о формате двоичных файлов Word, MS-DOC. Режим доступа: www.msdn.microsoft.com/ru-ru/library/office/gg615596(v=%20office.14).aspx (дата обращения 20.03.2015).
4. Object Linking and Embedding (OLE) Property Set Data Structures. Available at: www.msdn.microsoft.com/en-us/library/dd942421.aspx , accessed 21.03.2015.
5. Rich Text Format (RTF) Version 1.5 Specification. Available at: www.biblioscape.com/rtf15_spec.htm, accessed 23.03.2015.
6. Word Document (DOC) file format. Available at: www.forensicswiki.org/wiki/DOC, accessed 21.03.2015.
7. Computer Forensics Education – the Open Source Approach, Ewa Huebner, Derek Bem, and Hon Cheung, 2010. Available at: http://www.springer.com/978-1-4419-5802-0, accessed 25.03.2015.
8. Forensic and Anti-Forensic Techniques for OLE2-Formatted Documents, Jason M. Daniels, 2008. Available at: http://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1140&context=etd, accessed 27.03.2015.
9. Document Metadata and Computer Forensics, James Madison University Infosec Techreport, Department of Computer Science, 2006. Available at: http://www.infosec.jmu.edu/reports/jmu-infosec-tr-2006-003.pdf, accessed 02.04.2015.