DOI: 10.15514/ISPRAS-2021-33(4)-10

Алгоритм маркирования текстовых документов на основе изменения интервалов между словами, обеспечивающий устойчивость к преобразованию формата

[1] *А.В. Козачок, ORCID: 0000-0002-6501-2008<a.kozachok@academ.msk.rsnet.ru>*
[1] *С.А. Копылов, ORCID: 0000-0003-2841-5243 <gremlin.kop@mail.ru>*
[1] *П.Н. Горбачев, ORCID: 0000-0002-4511-0348 <pavel.gorbachev@list.ru>*
[2] *А.Е. Гайнов, ORCID: 0000-0003-3887-5374 <gae@mil.ru>*
[2] *Б.В. Кондратьев, ORCID: 0000-0003-0510-651X <gae@mil.ru>*

[1] *Академия Федеральной службы охраны Российской Федерации, 302015, Россия, г. Орёл, ул. Приборостроительная, д. 35*
[2] *Министерство обороны Российской Федерации, 119160, г. Москва, ул. Знаменка, д.19*

Аннотация. В статье представлен алгоритм маркирования электронных текстовых документов, основанный на внедрении идентификационной информации за счет изменения величин интервалов между словами. Разработка алгоритма направлена на повышение защищенности документов, содержащих текстовую информацию, от утечки по каналу, обусловленному передачей напечатанных на бумаге документов, а также соответствующих электронных копий бумажных документов. В процессе разработки алгоритма маркирования проведен анализ существующих средств защиты бумажных документов от утечки, рассмотрены практические решения в области защиты текстовых документов, определены их достоинства и недостатки. В качестве подхода к внедрению информации выступает алгоритм изменения величин интервалов между словами. Изменение величин интервалов между словами основано на встраивании удлиненного пробела в выделенные области строк текста и корректировке остальных значений величин интервалов между словами на рассчитанные значения. Для обеспечения инвариантности встроенного маркера к печати и последующему сканированию или фотографированию разработаны алгоритмы формирования областей встраивания и матрицы встраивания. В процессе формирования областей встраивания из строк текста исходного документа формируются массивы пробелов, состоящие из пар: по четыре и два пробела или по два пробела. Посредством встраиваемой информации в сформированных областях определятся места встраивания удлиненного пробела. В процессе встраивания маркера формируется матрица встраивания, содержащая значения смещения слов, и осуществляется ее встраивание в исходный документ при печати. Применение разработанного алгоритма маркирования позволяет внедрять в структуру текста электронного документа маркер, инвариантный к преобразованию формата электронного документа в бумажный посредством печати и обратно через сканирование или фотографирование. Представлены особенности и ограничения разработанного алгоритма маркирования. Определены направления дальнейших исследований.

Ключевые слова: защита от утечки информации; маркирование; распознавание образов; обработка изображений; текстовые документы

Для цитирования: Козачок А.В., Копылов С.А., Горбачев П.Н., Гайнов А.Е., Кондратьев Б.В. Алгоритм маркирования текстовых документов на основе изменении интервала между словами, обеспечивающий устойчивость к преобразованию формата. Труды ИСП РАН, том 33, вып. 4, 2021 г., стр. 131-146. DOI: 10.15514/ISPRAS–2021–33(4)–10

Text documents marking algorithm based on interword distances shifting invariant to format conversion

[1] *A.V. Kozachok, ORCID: 0000-0002-6501-2008 <a.kozachok@academ.msk.rsnet.ru>*
[1] *S.A. Kopylov, ORCID: 0000-0003-2841-5243 <gremlin.kop@mail.ru>*
[1] *P.N. Gorbachev, ORCID: 0000-0002-4511-0348 <pavel.gorbachev@list.ru>*
[2] *A.E. Gaynov, ORCID: 0000-0003-3887-5374 <gae@mil.ru>*
[2] *B.V. Kondrat'ev, ORCID: 0000-0003-0510-651X <gae@mil.ru>*

[1] *Academy of Federal Guard Service, 35, Priborostroitel'naya st., Orel, 302015, Russia.*
[2] *Ministry of Defence of the Russian Federation, 19, Znamenka st., Moscow, 119160, Russia*

Abstract. The article presents an electronic text documents marking algorithm based on the identification information embedding by changing the values of the intervals between words (interwords distance shifting). The algorithm development is aimed at increasing the documents containing text information security from leakage through the channel due to the transfer of documents printed on paper, as well as the corresponding electronic copies of paper documents. In the marking algorithm developing process, an existing tools analysis of protecting paper documents from leakage was carried out, practical solutions in the field of protecting text documents were considered, their advantages and disadvantages were determined. The interwods distance shifting algorithm acts as an approach to the information embedding in electronic documents. Changing the values of interwords distance is based on embedding the normalized space in the selected areas of text lines and adjusting the remaining values of the spacing between words by the calculated values. To invariance ensure of the embedded marker for printing and subsequent scanning or photographing, formation algorithms of embedding regions and embedding matrix have been developed. In the embedding regions forming process from the text lines of the source document, arrays of spaces are formed, consisting of pairs: four and two spaces or two spaces. By means of the embedded information in the formed areas, the places where the normalized space is inserted is determined. In the embedding a marker process, an embedding matrix is formed, containing the values of the word displacement, and it is embedded in the original document in the process of printing. The developed marking algorithm usage makes it possible to introduce a marker into the electronic document text structure that is invariant to the format transformation of an electronic document into a paper one and vice versa. In addition, the developed marking algorithm features and limitations are presented. Directions for further research identified.

Keywords: information leakage protection; marking; pattern recognition; image processing; text documents

For citation: Kozachok A.V., Kopylov S.A., Gorbachev P.N., Gaynov A.E., Kondrat'ev B.V. Text documents marking algorithm based on interword distances shifting invariant to format conversion. Trudy ISP RAN/Proc. ISP RAS, vol. 33, issue 4, 2021. pp. 131-146 (in Russian). DOI: 10.15514/ISPRAS–2021–33(4)–10

1. Введение

Совершенствование средств защиты информации является одним из наиболее актуальных направлений развития подходов к обеспечению информационной безопасности. Стремительный переход всех сфер общественных отношений в информационное пространство позволил не только повысить степень автоматизации предоставляемых услуг и сервисов конечным пользователям, но и выявить существующие ограничения и дефекты, присущие средствам защиты.

Анализ инцидентов информационной безопасности, связанных с утечкой конфиденциальной информации и персональных данных показал, что за 2020 год в мире зарегистрировано 2395 случаев утечки информации ограниченного доступа, при этом на Россию приходится примерно 17 процентов (404 случая утечки) от общего числа [1, 2]. По типу нарушителя наибольшее число утечек осуществлено за счет действий внутренних нарушителей (79 %), которые в большинстве случаев осуществлены умышлено. Наибольшее распространение среди каналов утечки получили: сетевые (61%), службы мгновенных сообщений,

осуществляющих передачу текстовых, голосовых и видео сообщений (19,7%), а также канал утечки бумажных документов (15%). Реализация столь значительного числа нарушений информационной безопасности стала возможна ввиду ориентированности существующих средств защиты на противодействие вредоносным и противоправным действиям внешних нарушителей.

Для защиты от утечек широкое распространение получили средства межсетевого экранирования и системы предотвращения утечек информации (Data Loss/Leak Prevention, DLP-системы). Межсетевой экран (брандмауэр, firewall) – аппаратное и\или программное средство, осуществляющее анализ, мониторинг и фильтрацию проходящего сетевого трафика согласно установленных правил безопасности [3–5]. Применение методов блокировки и анализа передаваемого трафика на разных уровнях модели OSI средствами межсетевого экранирования позволяет снизить количество утечек, осуществляемых по сетевым каналам, а также по каналам утечки, использующим службы мгновенных сообщений, осуществляющих передачу текстовых, голосовых и видео сообщений. В то же время указанные средства защиты не способны противодействовать утечкам конфиденциальной информации по каналу, обусловленному преобразованием формата электронного документа, содержащего текстовую информацию, в бумажный документ посредством печати и обратно – в цифровое изображение посредством сканирования распечатанного документа.

Для устранения указанного недостатка могут быть применены DLP-системы, осуществляющие идентификацию конфиденциальных данных, направляемых в другие сети, находящиеся в распределенных репозиториях (банках данных), а также хранящихся у конечного пользователя [6–9]. В зависимости от используемого метода защиты DLP-системы могут быть разделены на: сигнатурные методы (технология цифровых отпечатков пальцев), методы поиска по словарю, а также методы маркирования. Применение технологии цифровых отпечатков пальцев позволяет добиться высокой точности обнаружения в исследуемом трафике конфиденциальной информации. К недостаткам данной технологии относится отсутствие стойкости сигнатуры (шаблона) к применению преобразований и внесению искажений (в том числе и преднамеренных), а также невозможность выделения (формирования) сигнатуры для некоторых типов данных. Кроме того, сигнатурные методы и методы поиска по словарю накладывают на процесс обнаружения требование по созданию и ведению актуальной базы сигнатур или шаблонов.

В отличие от технологии цифровых отпечатков пальцев, в процессе использования технологии маркирования осуществляется встраивание маркера (идентификационной информации) в исходный документ [10]. Наличие встроенного маркера позволяет повысить стойкость защищаемой информации к осуществлению различных преобразований и внесению искажений. Указанная особенность позволяет использовать методы маркирования для повышения защищенности бумажных документов от утечки.

Анализ существующих решений в области маркирования документов, выводимых на печать, показал, что для защиты бумажных документов от утечки существуют следующие программные продукты Trace Doc [11], EveryTag [12] и SafeCopy [13]. Однако существенным недостатком всех продуктов является необходимость создания защищенной базы данных, содержащей или оригиналы электронных текстовых документов, подготовленных к печати, или их подписанные копии. Наличие указанного недостатка не позволяет предотвратить утечку бумажного документа, не имеющего подписанного электронного оригинала.

Для устранения указанного недостатка и повышения защищенности бумажных документов, а также соответствующих электронных копий от утечки необходимо разработать алгоритм маркирования текстовых документов, подготовленных к печати, обеспечивающий инвариантность встроенной информации к осуществлению преобразования формата и возможность извлечения встроенного маркера только из подписанного документа.

2. Алгоритм маркирования текстовых документов на основе изменения интервалов между словами

В качестве предложенного подхода к маркированию электронных документов, содержащих текстовую информацию, в процессе печати выступает подход к внедрению информации в структуру и формат текста за счет горизонтального сдвига слова внутри строки [14].

Подход к маркированию электронных документов, основанный на изменении положения (горизонтального сдвига, изменения величины интервала между словами) слов внутри текстовых строк, используется в работах Хуанга (Ding Huang) и др. [15], Алаттара (Adnan M. Alattar) и др. [16], Яна (Huijuan Yang) и др. [17].

В работе [15] в качестве электронного документа, в который осуществляется встраивание, выступает электронное текстовое изображение (изображение, содержащее исходный текстовый документ). Встраивание информации реализуется посредством изменения расстояния между словами и длин каждого слова таким образом, чтобы среднее расстояние между словами в каждой строке представляло собой синусоидальную волну определенной фазы и частоты. К достоинствам предложенного подхода относятся: возможность извлечения встроенной информации только из подписанных изображений, стойкость к осуществлению преобразования "печать-сканирование", фильтрация изображения и поворот на небольшие углы. Существенным недостатком предложенной схемы маркирования является возможность внедрения информации только в тексты, выравненные по ширине, а также требование по наличию информации о фазе и частоте синусоидальной волны на этапе извлечения встроенной информации.

По аналогии с исследованием [16], в работе Алаттара [15] представлена схема маркирования электронных текстовых документов за счет изменения величин интервалов между словами, позволяющая осуществлять извлечение встроенной информации, как из электронных документов, так и из соответствующих им бумажных копий. В процессе встраивания информации осуществляется увеличение или уменьшение величины интервала между словами на установленное значение в зависимости от значения символа встраиваемой последовательности. Предложенная схема позволяет повысить величину предельно достижимой емкости встраивания до 300 бит на страницу текста формата Letter. В свою очередь, процесс извлечения характеризуется наличием "пачек" ошибок (последовательно идущих ошибок), что накладывает требование по использованию помехоустойчивых кодов, обладающих свойством перемежения. Применение подобных кодов, в свою очередь, снижает размер встраиваемой информации.

В отличие от исследований [15, 16], в работе [17] предложена схема изменения величин интервалов между словами, основанная на измерении не только интервалов между словами, но и интервалов между символами на краях слов. В процессе внедрения осуществляется детектирование смежных друг с другом слов, а также интервалов, как между словами, так и между соседними символами, находящихся на краях данных слов. Посредством полученных значений осуществляется увеличение или уменьшение интервала между словами на установленное значение. Разработанный алгоритм позволяет осуществлять встраивание и извлечение встроенной информации из текстовых изображений и характеризуется высокой точностью извлечения. При этом в случае извлечения встроенной информации из отсканированных или сфотографированных изображений возрастает количество ошибок, связанных с ошибочным детектированием положения краевых символов. Указанные ошибки возникают из-за объединения соседних слов или слова и предлога воедино.

Проведенный анализ разработанных подходов к маркированию электронных документов, основанных на изменении интервалов межу словами, позволяет сделать вывод о наличии ограничений в процессе защиты бумажных документов от утечки, обусловленной преобразованием формата бумажного документа в электронный вид посредством применения операции "печать-сканирование" или "печать-фотографирование". В связи, с чем

для повышения точности извлечения и невидимости встроенных данных предложена схема маркирования $4BLK+2$. В качестве областей встраивания выступают последовательности из четырех или двух последовательно идущих пробелов. Встраивание информации осуществляется за счет изменения величины пробелов между словами следующим образом: последовательности из четырех пробелов соответствует два символа встраиваемой информации, двум пробелам – один символ. Применение данной схемы позволяет учесть недостатки подходов маркирования, описанных в работах [15–16].

Разработанный алгоритм маркирования электронных документов представлен на рис. 1.

Data: Документ D_O, встраиваемая информация L
Result: Подписанный документ D_M
$Im \leftarrow$ **DocumentConverting** (D_O)
$Lines \leftarrow$ **TextDetection** (Im)
$N \leftarrow |Lines|$
for $l \leftarrow 1$ **to** N **do**
  if $(|Lines[l]| > 1)$ **then**
    $SP[l] \leftarrow$ **CalcSpaceInLine**$(Lines[l])$
    $Long[l] \leftarrow$ **LongSpace**$(\frac{\sum SP[l]}{|SP[l]|})$
    $BLK = \max(\{bl \in \overline{0,|SP[l]|} : bl \mod 4 \equiv 0\})$
    $\{CS^1[l]\}, CS^2[l] \leftarrow$ **DevideSpaces**$(SP[l])$,
    $\forall cs_a \in \bigcup CS^1[l]\ \exists ss_b \in SP[l]: cs_a \equiv ss_b, a = b, b \leqslant BLK$
    $\forall cs_a \in CS^2[l]\ \exists ss_b \in SP[l]: cs_a \equiv ss_b, a = b, b > BLK, b \leqslant BLK+2$
    $\forall cs \in CS^4[l]: |cs| = 4$
    $|CS^2[l]| \subset \{0,2\}, |\bigcup CS^4[l] \cup CS^2[l]| \subset \{|SP[l]|, |SP[l]|-1\}$
  else
    $SP[l], CS^4[l], CS^2[l] \leftarrow \varnothing$
$E \leftarrow$ **ReedSolomon** (L)
$m \leftarrow 0$
for $l \leftarrow 1$ **to** N **do**
  for $k \leftarrow 1$ **to** $|CS^4[l]|$ **do**
    if $(CS^4[l][k] \neq \varnothing)$ **then**
      switch $(E_{\mathbf{GF}(m,|E|)}, E_{\mathbf{GF}(m+1,|E|)})$ **do**
        case 0, 0 **do**
          $CS^4[l][k] \leftarrow [Long[l], 0, 0, 0]$, $m = m+2$
        case 0, 1 **do**
          $CS^4[l][k] \leftarrow [0, Long[l], 0, 0]$, $m = m+2$
        case 1, 0 **do**
          $CS^4[l][k] \leftarrow [0, 0, Long[l], 0]$, $m = m+2$
        otherwise do
          $CS^4[l][k] \leftarrow [0, 0, 0, Long[l]]$, $m = m+2$
  if $(CS^2[l] \neq \varnothing)$ **then**
    if $(E_{\mathbf{GF}(m,|E|)} == 0)$ **then**
      $CS^2[l] \leftarrow [Long[l], 0]$, $m{+}{+}$
    else
      $CS^2[l] \leftarrow [0, Long[l]]$, $m{+}{+}$
  $av[l] \leftarrow 0$
  if $(SP[l] \neq \varnothing)$ **then**
    $av[l] \leftarrow \frac{\sum SP[l] - \sum\sum SP^4[l] - \sum SP^2[l]}{|SP[l]|}$
  $MLines[l] \leftarrow (CS^4[l], CS^2[l], av[l])$
$D_m \leftarrow$ **ChangeSpaces**$(D_O, Lines, MLines)$
return D_M

Рис. 1. Алгоритм маркирования текстовых электронных документов, основанный на изменении интервалов между словами
Fig. 1. Electronic text document marking algorithm based on interword distances shifting

Применение указанного подхода позволяет осуществить перцептивно невидимое внедрение информации, не внося изменений в структуру, семантику и синтаксис исходного документа. Кроме того, алгоритмы данной группы характеризуются высокой невидимостью встроенной

информации к визуальному анализу, а также емкостью встраивания достаточной для внедрения 50-200 бит информации.

Алгоритм маркирования электронных документов, содержащих текстовую информацию – электронных текстовых документов, подготовленных к печати, состоит из:

- преобразования формата электронного текстового документа в изображение;
- выделения строк текста из текстовых областей изображения;
- выделения областей встраивания;
- кодирования встраиваемой информации;
- формирования матрицы встраивания на основе закодированной информации и областей встраивания;
- смещения областей изображение соответствующих словам в выделенных текстовых строках согласно рассчитанной матрице встраивания.

На первом этапе (шаг 1 алгоритма маркирования) электронный текстовый документ D_O подвергается конвертации формата в изображение Im. Под электронным текстовым документом в контексте работы понимается документ, содержащий сплошной текст, текст, разбитый на графы, а также другую текстовую информацию, представленную в электронной форме [18, 19].

На втором этапе (шаг 2 алгоритма маркирования) осуществляется формирование текстовых областей изображения и выделение строк текста *Lines* посредством функции **TextDetection**.

2.1 Выделение строк текста в текстовых областях изображения

Алгоритм выделения строк текста *Lines*, реализуемый функцией **TextDetection**, представлен на рис. 2.

Function TextDetection(Im)
$Im_{rast} \leftarrow$ **Rasterization**(Im)
$k = 1$
if $(width(Im_{rast}) < 1654)$ **then**
  $k \leftarrow \frac{1654}{width(Im_{rast})}$
  $Im_{rast} =$ **Resize**(Im_{rast}, k)
if $(width(Im_{rast}) > 9920)$ **then**
  $k \leftarrow \frac{9920}{width(Im_{rast})}$
  $Im_{rast} \leftarrow$ **Resize**(Im_{rast}, k)
for $i \leftarrow 0$ **to** 4 **do**
  $Im_{filt} \leftarrow$ **BlurGaussianFilter**$(Im_{rast}, [5 \times 5])$
$Im_{filt} \leftarrow$ **Dilate**$(Im_{filt}, [3 \times 3])$
$Im_{mono} \leftarrow$ **Binarization**(Im_{filt})
$Areas \leftarrow$ **DetectConnectedAreas**(Im_{mono})
$Links \leftarrow$ **DetectAreaLink**$(Areas)$
$RawLines \leftarrow$ **DetectLines**$(Links)$
for $l \leftarrow 1$ **to** $|RawLines|$ **do**
  $Lines[i] \leftarrow$ **SplitToWords**$(RawLines[i])$
return $Lines$

Рис. 2. Функция выделение строк текста ***TextDetection***
Fig. 2. Text lines detection function ***TextDetection***

В процессе извлечения строк текста *Lines* из изображения функцией **TextDetection** осуществляется детектирование текстовых областей. Из полученных текстовых областей формируются текстовые последовательности, составляющие строки текста исходного

документа D_O . На шаге 1 осуществляется растеризация векторных элементов изображения Im, при их наличии в соответствии с заданным разрешением печати.

Растеризованное изображение Im_{rast} подвергается процедуре нормализации изображения (шаги 2-8). В случае если ширина $width(Im_{rast})$ превышает значение в 9920, то $width$ уменьшается до значения 9920, в случае если $width(Im_{rast})$ меньше 1654, то $width$ увеличивается до этого значения. В результате нормализации формируется нормализованное растрированное изображение Im_{rast} (рис. 3а). Процедура нормализации разрешения, с одной стороны, является оптимизацией, поскольку задачи детектирования текстовых областей не требуют сверхвысокого разрешения изображения, но в то же время сложность таких методов растет нелинейно, и работа над высокодетализированным изображением приведёт к взрывному росту вычислительной сложности. С другой стороны, ограничение возможного разрешения изображения некоторым диапазоном значений позволяет на последующих шагах заложить эвристики и фиксированные значения матриц, используемых в операциях над изображением.

Рис. 3. Результаты выделения строк из изображения
Fig. 3. Lines detection results from image

Нормализованное изображение Im_{rast} подвергается фильтрации, включающую гауссовский фильтра размытия (шаги 10-11) и операцию растягивания светлых областей (шаг 12). Функция **BlurGaussianFilter** фильтрует изображениеIm_{rast} гауссовским Фильтром размытия матрицей свертки 5×5 (рис 3б). Функция **Dilate** осуществляет растягивание (расширение) светлых областей изображения Im_{filt}, используя матрицу свертки 3×3 (рис 3в). Фильтрация изображения позволяет снизить шум и повысить точность детектирования текстовых областей. Отфильтрованное изображение Im_{filt} подвергается бинаризации – формированию черно-белого изображения Im_{mono} посредством функции **Binarization** (рис. 3г). Совокупность такой последовательности трех операций так же приводит к уменьшению различий между бинаризиноваными изображениями одного и того же документа, полученных из разных источников. Фактическим результатом являются остовы текстовых символов, которые необходимо объединить в совокупности, принадлежащие к одной строке, одному слову.

Выделение множества смежных областей одного цвета в изображении Im_{mono} реализовано процедурой **DetectConnectedAreas** (шаг 14), путём поэтапного обхода всех пикселей изображения с маской отсечения для уже посещенных координат. На выходе формируется система множеств, элементы которой соответствуют координатам пикселей, формирующих остовы символов. При этом, в силу особенностей языковой записи, шрифта и искажений изображения, некоторые символы могут распознаваться как последовательность из более чем одной области, например, символы "ы", ";", "й". И наоборот, когда одной области могут принадлежать последовательности из нескольких символов, к примеру, сливающиеся декоративные элементы последовательности символов «ны» для шрифтов с засечками. В процедуре **DetectConnectedAreas** реализована эвристика отбрасывания тех элементов системы множеств, которые имеют слишком малую мощность для распознаваемого

изображения, что позволяет качественно отфильтровывать лишние шумы (рис. 3г, 3д). С теоретической точки зрения такая эвристика обоснована минимальным читаемым разрешением, поскольку совокупность из 2-3 пикселей шума, оставшаяся после этапов фильтрации изображения, не может представлять собой часть изображения пригодную для интерпретации человеком в качестве символа.

Рассчитанная система множеств, передается в процедуру **DetectAreaLink**, где с каждым элементом сопоставляются его базовые характеристики, включая центр области (центроид), высоту, ширину. Затем, для близких пар символов рассчитываться информация об их связанности, включая расстояние между ними, соотношение их сторон относительно друг друга. Текстовая область представляет собой совокупность связанных контуром отдельных символов, принадлежащих одной прямой, проходящей через центр символьных контуров. Для каждого символьного контура посредством функции **DetectAreaLink** рассчитывается его центр (центроид), который соединяется с центрами соседних символьных контуров, принадлежащих одной прямой линии (рис 3е). Символьные контуры, сформированные на шаге 15, представляют собой связанные области, из которых функцией **DetectLines** формируются строки текста *RawLines* (рис 3ж). На шагах 17-18 каждая текстовая строка разделяется на слова (рис. 3з). Полученный массив текстовых строк *Lines* представляет собой результат выполнения функции **TextDetection**, на основании которого формируются области встраивания и встраиваемая последовательность сдвигов слов в тексте.

2.2 Формирование областей встраивания

На третьем этапе (шаги 3-15 алгоритма маркирования) из массива строк текста *Lines* формируются области встраивания исходного документа D_O. Область встраивания представляет собой систему множеств, элементы которой эквивалентны словам строки текста и, в свою очередь, состоят из наборов координат пикселей, образующих эти слова. Предлагаемый алгоритм требует не менее трех слов в каждой кодируемой строке, поскольку крайние слова остаются на своих позициях для сохранения выравнивания текста и минимизации визуальных искажений. Таким образом, строки из одного-двух при встраивании метки остаются без изменений и игнорируются при её извлечении.

В процессе коррекции искажений требуется выполнить сдвиг слов внутри строки, обеспечивающий изменение соотношения длин обрамляющих их пробелов, при неизменности суммы всей совокупности, что создаёт особые требования к таковой операции. Незначительное изменение пробелов между словами может оказаться недостаточным для их различия на этапе извлечения информации, в то же время использование чрезмерно длинных пробелов, потребует равного их длине сокращения остальных. В свою очередь, указанная особенность может привести к недостаточности длины пробела для разделения двух и более слов и на этапе детектирования, и они будут обнаружены как одно.

Для решения этой проблемы была предложена схема кодирования "$4BLK+2$". В соответствии с этой схемой пробелы в строке, начиная с самого первого последовательно объединяться в BLK (шаг 8) блоков CS^4 по четыре пробела в каждом (шаг 9). При наличии в строке остатка, из двух-трех пробелов, не вошедших в блоки CS^4, первая пара формирует один блок CS^2. Для примера, строки из трех, четырех слов формирует $BLK = 0$ блоков CS^4 и один блок CS^2, пять-шесть слов дают $BLK = 1$ блок CS^4 и ноль блоков CS^2 и так далее. Формальные условия, которым должна удовлетворять формируемая совокупность $\{CS^4[l], CS^2[l]\}$, определены в строках 10-12. В такой схеме, для обычного текста имеется существенное преобладание блоков типа CS^4, что позволяет формировать сдвиги с одним удлиненным (удлиняющимся) пробелом, при вычитании этого удлинения из трех укорачивающихся. Подобный подход уменьшает ошибки в ходе детектирования текстовых областей с чрезмерно малыми пробелами, а также связывает схему кодирования с кодовой

структурой с четным числом блоков, снижая вероятность влияния ошибок вставки или выпадения одиночного пробела в строке.

В результате на третьем этапе алгоритма маркирования для каждой строки текста l формируется область встраивания – массив значений, состоящий из следующих элементов:

- $SP[l]$ – последовательность интервалов между словами (пробелов);
- $\{CS^4[l], CS^2[l]\}$ – область встраивания, состоящая из последовательностей по 4 или 2 пробела;
- $Long[l]$ – значение удлиненного пробела.

На четвертом этапе (шаг 16 алгоритма маркирования) осуществляется помехоустойчивое кодирование встраиваемой информации L в качестве маркера E. Использование помехоустойчивого кода позволяет компенсировать ошибки, возникающие в результате искажений маркируемого документа: в результате его подписи, добавления в отпечатанный документ печатей, утраты его части или ошибок в работе детектора текста.

2.3 Особенности формирования встраиваемого маркера

В качестве встраиваемой информации L может выступать следующая информация:

- идентификатор пользователя L_1;
- метка конфиденциальности документа L_2;
- метка, содержащая информацию о времени встраивания L_3;
- идентификатор средства вычислительной техники L_4;
- информация о целостности документа L_5.

Встраиваемая метка зависит структуры и размера встраиваемой информации L. Исходя из цели разработанного алгоритма маркирования – повышение защищенности конфиденциальной информации и персональных данных от утечек по каналу утечкой бумажных документов, структура встраиваемой информации L имеет вид: $L = L_1, L_2, L_3$.

Размер встраиваемой метки зависит от емкости встраивания электронного документа D_O и не может превышать величину предельно достижимой емкости встраивания. Под емкостью встраивания понимается то количество информации, которое может быть встроено в электронный документ выбранным подходом к внедрению информации. Для электронного текстового документа D_O величина предельно равна значению $|SP[l]|, \forall l \in 1 \dots N$ и определяется количеством интервалов между словами (числом пробелов) в тексте.

Исходя из особенностей электронных документов, которые могут быть заполнены текстом полностью, наполовину или могут содержать всего несколько строк текста, целесообразно ограничить размер встраиваемого маркера 48 битами. Учитывая тот факт, что в разработанном алгоритме маркирования для внедрения 1 бита информации необходимо два интервала между словами, то для внедрения 48 бит информации требуется не менее 96 интервалов между словами (пробелов) в тексте и соответствует документу, содержащему порядка 100-120 слов.

Для обеспечения стойкости встраиваемого маркера к осуществлению преобразования формата электронного текстового документа в бумажный вид, а также повышения точности извлечения встроенного маркера на шаге 16 алгоритма маркирования происходит помехоустойчивое кодирование встраиваемой информации L кодом Рида-Соломона. Применение указанного кода к последовательности маркера E размером 48 бит снижает информационную часть до значения в 32 бита, остальные 16 бит приходятся на проверочную часть.

Для формирования маркера E исходная информация L кодируется в двоичную систему счисления, при этом под каждый компонент L отводится определенное число бит: L_1 – 14 бит, L_2, – 4 бита, L_3 – 14 бит. Полученная последовательность кодируется кодом Рида-Соломона и передается на этап формирования встраиваемой последовательности.

2.4 Формирование матрицы встраивания

На пятом этапе (шаги 17-38 алгоритма маркирования) посредством закодированной информации E и областей встраивания ($SP[l]$, $\{CS^4[l], CS^2[l]\}$, $Long[l]$) осуществляется формирование матрицы встраивания *MLines*, содержащей позиции интервалов между словами и соответствующие значениям в пикселях. Для формирования последовательности *MLines* осуществляется последовательное считывание элементов областей встраивания $\{CS^4[l], CS^2[l]\}$. Для блока из четырех пробелов $CS^4[l]$ из закодированной последовательности E извлекается очередная пара бит, определяющая номер нормального пробела $Long[l]$ в блоке $CS^4[l]$. В зависимости от последовательно считанных символов $E_{GF(m,|E|}, E_{GF(m+1,|E|}$ $CS^4[l]$ формируется следующим образом:

- если два первых символа последовательности E имеют значение (0,0), то первый пробел в последовательности $CS^4[l]$ заменяется на нормированный пробел $Long[l]$ строки l, остальные элементы $CS^4[l]$ остаются неизменными;
- если два первых символа последовательности E имеют значение (0,1), то значение $Long[l]$ присваивается второму элементу, при этом остальные элементы $CS^4[l]$ остаются без изменений;
- если два первых символа последовательности E имеют значение (1,0), то значение $Long[l]$ присваивается третьему элементу, при неизменности остальных элементов $CS^4[l]$;
- в остальных случаях – только четвертый элемент $CS^4[l]$ заменяется на $Long[l]$;
- после любой из замен осуществляется переход к следующему символу кодовой последовательности E.

Для блоков из двух бит $CS^2[l]$ матрицы встраивания $\{CS^4[l], CS^2[l]\}$, в зависимости от символа $E_{GF(m,|E|}$, пробел $Long[l]$ подставляется следующим образом:

- если символ последовательности E имеет значение (0), то значение $Long[l]$присваивается первому элементу, при этом второй элемент $CS^2[l]$ не изменяется;
- в противном случае – второй элемент $CS^2[l]$ заменяется на $Long[l]$, при отсутствии изменений в первом;
- после любой из замен осуществляется переход к следующему символу кодовой последовательности E.

Стоит отметить, что в случае наложения сформированной матрицы встраивания на исходный текст без осуществления компенсации остальных пробелов в строке возможно появление артефактов вида: выход строки текста за поля, искажение структуры текста, сдвиг знаков пунктуации и прочие изменения в структуре исходного текста.

Наличие указанных искажений делает встроенный маркер заметным для визуального анализа и нарушает структуру текста. Для устранения указанного недостатка используется процедура компенсации пробелов в строке. На шагах 35-37 для каждой строки рассчитываются размеры всех остальных пробелов $av[l]$. Перерасчет размеров пробела позволяет добиться того, что все нормированные пробелы приняли одинаковое значение, а сумма всех пробелов в строке осталась неизменной относительно исходного документа.

Полученные значения пробелов $av[l]$ и массив значений пробелов областей встраивания $\{CS^4[l], CS^2[l]\}$ представляют собой матрицу встраивания $MLines$. Полученная матрица

используется на этапе встраивания встраиваемой информации в величины интервалов между словами функцией **ChangeSpaces**.

2.5 Внедрение матрицы встраивания в текстовые области исходного электронного документа

На заключительном этапе алгоритма маркирования осуществляется внедрение встраиваемой информации в электронный текстовый документ посредством изменения величин интервалов между словами (значений пробелов) исходного текстового документа D_O. При этом все пробелы в тексте электронного документа заменяются по следующему правилу:

- пробелы в строке l, соответствующие позициям нормального пробела $Long[l]$ в областям встраивания $\{CS^4[l], CS^2[l]\}$, заменяются на значение равное $Long[l]$;
- значение остальных пробелов строки l корректируются (уменьшаются или увеличиваются) до значения равного $av[l]$.

Изменение величин пробелов сформированной матрицы встраивания осуществляет функция **ChangeSpaces**, представленная на рис. 4.

```
   Function ChangeSpaces(D_O, Lines, MLines)
1  Src ← Rasterization(D_O)
2  Lines' ← LinkedForegroundObjects(Src, Lines)
3  Dst ← ClearForeground(Src, Lines')
4  fastMin ← 0
5  fastMax ← 0
6  for l ← 0 to |MLines| do
7      CS^4, CS^2, av ← MLines[l]
8      CS ← CS^4 ⋃ CS^2
9      SP ← CalcSpaceInLine(Lines[l])
10     cur ← 0
11     for k ← 1 to |CS| do
12         if (CS[k] == 0) then
13             CS[k] ← av
14         if (k > 0 and k < |CS| − 1) then
15             sp ← SP[k]
16             cur ← (CS[k − 1] − (sp − cur))
17             offset[k] ← Round(cur)
18         else
19             offset[k] ← 0
20         fast[offset[k]] ← fast[offset[k]] ⋃ GetWordPixelMask(Lines'[l], k)
21         fastMin ← min(fastMin, offset[k])
22         fastMax ← max(fastMax, offset[k])
23 for l ← fastMin to fastMax do
24     if (fast[l] ≠ ∅) then
25         mask ← [Height(Src) × Width(Src)]
26         for l ← fastMin to fastMax do
27             if (Width(fast[l][k]) + l < Width(Src)) then
28                 mask[fast[l][k]] ← 1
29         mask ← Dilate(mask, [5 × 5])
30         locs ← {pix ∈ mask : mask[pix] ≠ 0}
31         Dst[locs + l] ← Src[locs]
32     TD_M ← Dst
   return TD_M
```

Рис. 4. Функция изменения величин между словами ***ChangeSpaces***
Fig. 4. Interword distances shifting function ***ChangeSpaces***

В начале процесса внедрения матрицы встраивания проводится этап дополнительного детектирования шумов, на котором все объекты, не соотнесенные со строками детектором текста, связываются с ближайшим словом и включаются в его состав. На шаге 1 исходный текстовый документ D_O конвертируется в изображение и растеризуется. Из растеризованного

изображения Src и извлеченного на предыдущих этапах массива строк текста $Lines$ посредством функции **LinkedForegroundObjects** выделяется массив строк $Lines'$, содержащий объекты, не соотнесенные со строками $Lines$. Этот шаг позволяет соотнести со словами любые элементы изображения, которые на этапах детектирования текстовых областей могли быть отфильтрованы в качестве шума, и обеспечивает минимизацию визуальных искажений при изменении местоположения слов.

После этого на шаге 3 рассчитывается фоновое изображение, т.е. исходное множество пикселей Src приводится к такому виду, что значение цвета для всякого элемента множество Src, входящего в состав одного из подмножеств $Lines'$, замещается усреднением цвета ближайших к нему по матрице смежности элементов Src не принадлежащих какому либо подмножеству $Lines'$. Для каждой строки l из маски встраивания $Mlines$ выделяются значения CS^4, CS^2 и av. CS^4 и CS^2 формируют маску встраивания строки CS (шаг 8). Функция **CalcSpaceInLine** вычисляет значения величин интервалов между словами в массиве строк $Lines$ (шаг 9). Исходя из полученных значений областей встраивания, на шагах 11-19 осуществляется расчет величин интервалов между словами для каждой строки текста с учетом сформированной матрицы встраивания следующим образом:

- если в последовательности CS имеются еще не рассчитанные значения, то значениям интервалов между словами (пробелов) присваивается значение $av[l]$ (шаги 13-14);
- начиная со второго и заканчивая предпоследним словом в строке, осуществляется расчет смещения относительно исходной позиции слова *cur*, полученный результат округляется до целого функцией **Round** (шаги 15-20). В результате округления для каждой строки текста формируется массив смещений $offset$;
- во всех остальных случаях формируется нулевое смещение, что соответствуют формированию немаркированного документа D_O.

Из сформированного массива смещений и рассчитанной посредством функции **GetWordPixelMask** маски слов всех строк текста формируется маска заполнения, содержащая координаты формирующих слово пикселей и соответствующие им смещения. Кроме того, вычисляется наименьшее значение массива смещения $fastMin$ и наибольшее $fastMax$ (шаги 22 и 23 соответственно).

Для осуществления внедрения матрицы встраивания в текстовые области исходного электронного документа реализуется процедура морфологического изменения границы посредством использования ядра фильтра, кратного размеру изображения (шаги 25-30). После чего изображение заполняется значениями пикселей с координатами, сдвинутыми на рассчитанное значение смещения (шаг 31), включая нулевые смещения для первых и последних слов строк. В результате работы алгоритма маркирования формируется подписанный документ D_M, подготовленный к печати, т.е. документ, содержащий встраиваемую информацию L (маркер).

Предлагаемый на шагах 20-31 подход не только реализует разработанный алгоритм, но и обеспечивает оптимизацию, связанную с внедрением специальных переменных $fast$, $fastMin$, $fastMax$, позволяющих сгруппировать операции сдвига на идентичные значения, что актуально в виду их дискретности. Такая оптимизация уменьшает число дорогостоящих операций морфологического искажения (шаг 29), извлечения координат по маске (шаг 30), а также локализует операции доступа к памяти (шаг 31), увеличивая вероятность попадания в процессорные кэши.

Разработанный алгоритм маркирования, основанный на изменении величин интервалов между словами, позволяет осуществлять внедрение идентификационной информации в различные документы, содержащие текстовую информацию и подготовленные к печати. Стоит отметить, что, как и любой алгоритм маркирования, разработанный алгоритм имеет некоторые особенности.

3. Особенности разработанного алгоритма маркирования

Отличительная особенность разработанного алгоритма маркирования состоит в том, что алгоритм может быть применим только к документам, содержащим текст или текст, разбитый на графы, а также другую текстовую информацию. Помимо описанной, разработанный алгоритм характеризуется следующими особенностями [20, 21]:

- наличие стойкости встроенного маркера к осуществлению печати документа, а также применения операции сканирования или фотографирования к полученному бумажному документу;
- возможность маркирования любого используемого формата документа (.doc, docx, .odt, .pdf, .png, .jpg, и др.), а также отсутствие зависимости процесса внедрения от используемого текстового редактора (при работе с текстовыми документами);
- высокой емкостью встраивания информации, т. к. для встраивания 1 бита информации требуется только 2 интервала между словами (пробела) или 3 слова;
- отсутствие стойкости встроенного маркера к применению средств оптического распознавания символов.

Кроме рассмотренных особенностей необходимо описать процесс извлечения из бумажных документов встроенной информации. Извлечение встроенной информации реализуется аналогичным образом, что и встраивание. Стоит отметить, что извлечение встроенной информации реализовано для цифровых изображений, полученных посредством сканирования или фотографирования бумажных документов.

При этом применение указанных операций имеет следующие особенности:

- процесс сканирования характеризуется внесением в конечное изображение таких искажений как геометрические преобразования (поворот, масштабирование, сдвиг, обрезка и др.) и шумы сканера (дополнительные пиксели или области изображения, не присутствовавшие в бумажном документе);
- процесс фотографирования помимо внесения геометрических искажений в трех плоскостях (поворот, отклонение и наклон камеры) характеризуется различными видами искажений, вносимых объективом фотоаппаратуры и условиями съемки (дисторсия, прогрессирующее снижение освещенности, хроматические аберрации, муар и прочие).

Наличие этапа предварительной обработки в процессе выделения строк текста, реализуемого функцией **TextDetection**, позволяет устранить лишь некоторые из представленных искажений. Для устранения указанного недостатка необходимо разработать алгоритм извлечения встроенной информации из изображений соответствующим бумажным документам. Разработка алгоритма извлечения встроенной информации позволит осуществить количественную оценку точности извлечения, устойчивости встроенного маркера к осуществлению преобразований и внесению искажений и осуществить сравнительный анализ полученных результатов с аналогами.

4. Заключение

Предложенный подход к маркированию электронных документов, выводимых на печать, позволяет повысить защищенность бумажных документов, содержащих конфиденциальную информацию, а также персональные данные, в случае неконтролируемого распространения или передачи документов лицам, не имеющим к ним легитимного доступа. Кроме того, наличие встроенной информации позволяет однозначно установить факт утечки, а также идентифицировать нарушителя. Разработанный алгоритм может применяться в DLP-системах и системах расследования инцидента информационной безопасности. При этом

разработка дополнительных этапов процесса извлечения, направленных на снижение влияния процесса сканирования (фотографирования) на точность извлечения встроенной информации, а также экспериментальная оценка основных параметров алгоритма маркирования и сравнение полученных результатов с существующими аналогами является направлением дальнейших исследований.

Список литературы / References

[1]. Россия: утечки информации ограниченного доступа, 2020 год. InfoWatch. 2021, 30 стр. / Russia: Restricted Information Leaks, 2020. InfoWatch. 2021, 30 p. Available at: https://www.infowatch.ru/analytics/analitika/rossiya-utechki-informatsii-ogranichennogo-dostupa-2020-god, accessed 10.08.2021 (in Russian).

[2]. Исследование утечек информации ограниченного доступа в 2020 году. InfoWatch. 2021, 40 стр. / Research on restricted information leaks in 2020. InfoWatch. 2021, 40 p. Available at: https://www.infowatch.ru/analytics/analitika/issledovanie-utechek-informatsii-ogranichennogo-dostupa-v-2020-godu, accessed 10.08.2021 (in Russian).

[3]. Mukkamala P. P., Rajendran S. A survey on the different firewall technologies, International Journal of Engineering Applied Sciences and Technology, vol. 5, issue 1, 2020, pp 363-365.

[4]. Neupane K., Haddad R., Chen L. Next Generation Firewall for Network Security: A Survey. In Proc. of the SoutheastCon 2018, 2018, pp. 1-6.

[5]. Sharma R. K., Kalita H. K., Issac B. Different firewall techniques: A survey. In Proc. of the Fifth International Conference on Computing, Communications and Networking Technologies (ICCCNT). 2014, pp. 1-6.

[6]. Lopez G., Richardson N., Carvajal J. Methodology for Data Loss Prevention Technology Evaluation for Protecting Sensitive Information. Revista Politécnica, vol. 36, no. 3, 2015, pp. 60-69.

[7]. Alneyadi S., Sithirasenan E., Muthukkumarasamy V. A survey on data leakage prevention systems. Journal of Network and Computer Applications, vol. 62, 2016, pp. 137-152.

[8]. Jadhav P., Chawan P. M. Data Leak Prevention system: A Survey. International Research Journal of Engineering and Technology, vol. 6, no. 10, 2019, pp. 197-199.

[9]. Kozachok A.V., Kopylov S.A. et al. Text marking approach for data leakage prevention. Journal of Computer Virology and Hacking Techniques, vol. 15. no. 3, 2019, pp. 219-232.

[10]. Козачок А.В., Кпылов С.А. и др. Подход к извлечению робастного водяного знака из изображений, содержащих текст. Труды СПИИРАН, вып. 5(60), 2018 г., стр. 128-155 / Kozachok A.V., Kopylov S.A. et al. An Approach to a Robust Watermark Extraction from Images Containing Text. SPIIRAS Proceedings, issue 5(60), 2018, pp. 128-155 (in Russian).

[11]. Trace Doc. Available at: https://secretgroup.ru/trace-doc, Accessed 10.08.2021.

[12]. Unique Interface. EveryTag. Available at: https://everytag.ru/ui, Accessed 10.08.2021.

[13]. Safe Copy. Available at: https://www.niisokb.ru/products/safecopy, Accessed 10.08.2021.

[14]. Jalil Z., Mirza A.M. A Review of Digital Watermarking Techniques for Text Documents. In Proc. of the International Conference on Information and Multimedia Technology, 2009, pp. 230-234.

[15]. Huang D., Yan H. Interword distance changes represented by sine waves for watermarking text images, IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 12, 2001, pp. 1237-1245.

[16]. Alattar A.M., Alattar O.M. Watermarking electronic text documents containing justified paragraphs and irregular line spacing, In Proc. of the Conference on Security, Steganography, and Watermarking of Multimedia Contents, 2004, pp. 685-695.

[17]. Yang H., Kot A.C. Text document authentication by integrating inter character and word spaces watermarking, In Proc. of the IEEE International Conference on Multimedia and Expo (ICME), 2004, pp. 955-958.

[18]. Национальный стандарт Российской Федерации. Система стандартов по информации, библиотечному и издательскому делу. Делопроизводство и архивное дело. Термины и определения, ГОСТ Р 7.0.8–2013, Стандартинформ, 2013 г., 16 с. / National standard of the Russian Federation. System of standards on information, librarianship and publishing. Records management and organization of archives. Terms and difinitions, GOST R 7.0.8–2013, Standartinform, 2013, 16 p. (in Russian).

[19]. Национальный стандарт Российской Федерации. Единая система конструкторской документации. Общие требования к текстовым документам, ГОСТ Р 2.105–2019, Стандартинформ, 2019 г., 35 с. /

National standard of the Russian Federation. Unified system for design documentation. General requirements for textual documents, GOST R 2.105–2019, Standartinform, 2019, 35 p. (in Russian).

[20]. Библиотека маркирования текстовых документов при печати за счет горизонтального смещения слов, Свидетельство о государственной регистрации программ для ЭВМ № 2020667592 от 24.12.2020, Россия, заявка № 2020666902 от 17.12.12020 / Library for marking text documents when printing due to horizontal displacement of words, Certificate of state registration of computer programs № 2020667592 dated 12.24.2020, Russia, application № 2020666902 dated 17.12.12020 (in Russian).

[21]. Модуль маркирования текстовых документов при печати для ОС семейства Windows, Свидетельство о государственной регистрации программ для ЭВМ № 2020667579 от 24.12.2020, Россия, заявка № 2020666721 от 17.12.12020 / Module for marking text documents when printing for Windows family OS, Certificate of state registration of computer programs № 2020667579 dated 12.24.2020, Russia, application № 2020666721 dated 12.17.12020 (in Russian).

Информация об авторах / Information about authors

Александр Васильевич КОЗАЧОК – доктор технических наук, доцент, сотрудник Академии Федеральной службы охраны Российской Федерации. Его научные интересы включают: информационная безопасность, защита от несанкционированного доступа, математическая криптография, теоретические проблемы информатики.

Alexander Vasilievich KOZACHOK – Doctor of Technical Sciences, Associated Professor. Employer of the Academy of Federal Guard Service. His research interests include: information security, unauthorized access protection, mathematical cryptography and theoretical problems of computer science

Сергей Александрович КОПЫЛОВ является сотрудником Академии Федеральной службы охраны Российской Федерации. Его научные интересы включают: методы машинного обучения, обработка цифровых изображений, текстовая стеганография.

Sergey Alexandrovich KOPYLOV is an employer of the Academy of Federal Guard Service. His research interests include machine learning methods, digital image processing, text steganography.

Павел Николаевич ГОРБАЧЕВ является сотрудником Академии Федеральной службы охраны Российской Федерации. Его научные интересы включают: информационная безопасность, методы машинного обучения, распознавание образов, текстовая стеганография, обработка изображений.

Pavel Nikolaevich GORBACHEV is an employer of the Academy of Federal Guard Service. His research interests include: information security, machine learning methods, pattern recognition text steganography and image processing.

Артур Евгеньевич ГАЙНОВ является сотрудником Министерства обороны Российской Федерации. Сфера научных интересов: безопасность информации, защита информации от несанкционированного доступа, построение информационных систем в защищённом исполнении.

Artur Evgenevich GAYNOV is an employer of the Ministry of Defence of the Russian Federation. Research interests: information security, information unauthorized access protection, information systems construction in a secure design.

Борис Владимирович КОНДРАТЬЕВ является сотрудником Министерства обороны Российской Федерации. Сфера научных интересов: информационная безопасность, защита информации от утечки по техническим каналам, сертификация программного обеспечения по требованиям безопасности информации, методы сокрытия информации.

Boris Vladimirovich KONDRAT'EV is an employer of the Ministry of Defence of the Russian Federation. Research interests: information security, protection against information leakages through technical channels, software certification in accordance with information security requirements, information hiding methods.