CODIFICACIÓN AUTOMATIZADA DE EVENTOS A PARTIR DE TEXTO ESCRITO EN ESPAÑOL

Javier Osorio[1]
Cornell University

Mayo 2014

1. INTRODUCCIÓN

Los avances en el área de tecnologías de la información de los últimos años han abierto la posibilidad de generar, compartir, difundir, reproducir, acceder y almacenar volúmenes masivos de información digital a nivel global. La "revolución de datos" se refiere a la velocidad y frecuencia en la producción y transmisión de datos digitales que fluyen desde una gran variedad de fuentes de información. La disponibilidad de información ofrece un potencial sin precedentes para ampliar nuestro conocimiento acerca de una amplia gama de comportamientos sociales. No obstante, los grandes datos no son una panacea. Como bien lo señala la Organización de las Naciones Unidas, la clave para concretar dicho potencial radica en realizar un análisis sistemático y riguroso de datos que permita transformar grandes colecciones de información imperfecta, compleja, desordenada y desvinculada en conocimiento útil y accionable (United Nations Global Pulse 2012).

Posiblemente, uno de los sectores más activos en el análisis de grandes datos es el área referente a temas de seguridad. La importancia de identificar, prevenir y neutralizar posibles amenazas a la seguridad internacional y doméstica ha favorecido el desarrollo de tecnologías y metodologías para procesar grandes volúmenes de información y generar productos de análisis (O´Brien 2012, Subrahmanian 2013, Leetrau y Schrodt 2013). La realización de este potencial se concentrada principalmente en países desarrollados gracias a la creación de sistemas de procesamiento específicamente diseñados para analizar información escrita en idioma inglés. Sin embargo, este enfoque anglo-céntrico excluye una cantidad masiva de información generada en otros idiomas. Esta deficiencia es particularmente preocupante si consideramos que sólo el 8.2

[1] Javier Osorio es Doctor en Ciencia Política por la Universidad de Notre Dame. Actualmente es Investigador Asociado del Mario Einaudi Center for International Studies en la Universidad de Cornell. El desarrollo de EVENTUS ID fue posible gracias al generoso apoyo financiero del Doctoral Dissertation Research Improvement Grant de la National Science Foundation (award SES-1123572); la beca Drugs, Security and Democracy (DSD) del Social Science Research Council y el Open Society Foundations; el Programa de Beca Doctoral Jennings Randolph del United States Institute of Peace; y el Graduate Research Grant del Kellogg Institute for International Studies de la Universidad de Notre Dame. Los refinamientos finales del programa fueron posibles gracias al apoyo del Program on Order, Conflict and Violence de la Universidad de Yale; la Beca Doctoral de la Fundación Harry Frank Guggenheim; y la beca postdoctoral del Mario Einaudi Center for International Studies de Cornell University. Correspondencia con el autor: javier.osoriozago@gmail.com

por ciento de la población mundial es angloparlante, lo cual indica que la información producida por el resto de la población es ignorada (Lewis, Simons y Fenning 2013).

En el caso particular de los países de habla hispana, la falta de un sistema que permita analizar información escrita en idioma español implica ignorar efectivamente una cantidad masiva de datos relevantes, oportunos, detallados y de alta calidad. Esta limitación implica el desaprovechamiento de un gran volumen de información que podría ser utilizado para generar datos sistematizados, productos de análisis y conocimiento valioso en temas prioritarios para América Latina. Dicha carencia es particularmente lamentable en materia de seguridad. Dado el problema generalizado de inseguridad que asedia a la región, la falta de datos confiables y sistemáticos impide el desarrollo de diagnósticos acertados y la generación de recomendaciones responsables de política pública.

Con el objetivo de atender dichas deficiencias, el presente capítulo presenta EVENTUS ID, una herramienta específicamente diseñada para la codificación computarizada de eventos a partir de información escrita en idioma español. EVENTUS ID fue inicialmente desarrollado por Osorio y Reyes para identificar patrones y tendencias de violencia relacionada con el crimen organizado en México (Osorio 2013, Osorio y Reyes 2014). Sin embargo, el potencial de este programa se extiende mucho más allá del estudio del comportamiento criminal, ya que puede ser utilizado para codificar eventos de cualquier tipo con base en texto escrito en español. El contenido de este documento brinda orientación técnica para el uso de EVENTUS ID con la intención de que los usuarios desarrollen sus propios proyectos de codificación automatizada de eventos.

La estructura del capítulo está compuesta de siete secciones. La primera parte presenta el programa EVENTUS ID para codificar eventos de manera automatizada con base en texto escrito en español. Este apartado también discute las ventajas y limitaciones del procesamiento de lenguaje natural frente a esquemas tradicionales de codificación humana. El segundo apartado presenta el esquema general de codificación de eventos implementado por EVENTUS ID. La tercera parte señala las características y formato del texto a ser procesado por el programa. La cuarta sección indica los pasos a seguir para codificar eventos con base en el desarrollo de diccionarios de actores y verbos. Este apartado también describe el funcionamiento de los algoritmos que EVENTUS ID utiliza para identificar eventos. El quinto segmento discute el proceso de geo-referenciación de eventos, así como los diccionarios de lugares y el algoritmo de identificación de localidades. Posteriormente, el sexto apartado hace referencia a la validación y recodificación de eventos generados por el protocolo de codificación. Finalmente, la séptima sección muestra un ejemplo de aplicación de este programa para el análisis de patrones de violencia relacionada con el crimen organizado en México.

1.1 Presentación de EVENTUS ID

EVENTUS ID es un sistema supervisado que permite la codificación de eventos a partir de texto escrito en español. El programa es *supervisado,* ya que requiere la intervención humana para el desarrollo de los diccionarios usados como categorías de búsqueda en el proceso de codificación y para evaluar la precisión del proceso de codificación. El programa es capaz de realizar la *codificación de eventos* gracias a un sistema de reconocimiento de patrones que permite identificar información acerca de quién le hizo qué a quién. La flexibilidad de dos algoritmos de codificación de eventos aunados a un protocolo de identificación de lugares plenamente integrado al programa permite que EVENTUS ID detecte eventos a nivel sub-nacional. Este programa constituye el primer sistema de codificación de eventos específicamente diseñado para procesar *información escrita en idioma español*.

Inicialmente, EVENTUS ID fue desarrollado como parte de la investigación doctoral realizada por Osorio, titulada "Hobbes on Drugs: Understanding Drug Violence in México", la cual analiza las dinámicas de violencia criminal en dicho país (Osorio 2013). La codificación de eventos sirvió para identificar una extensa gama de acciones violentas y no violentas que utilizan las fuerzas de seguridad pública para combatir a las organizaciones criminales, así como un amplio menú de tácticas violentas perpetradas por grupos criminales hacia las autoridades y en contra de grupos rivales. En un inicio, el proyecto de investigación contempló utilizar TABARI, un programa de codificación de eventos desarrollado por Schrodt (2014). Sin embargo, TABARI fue desarrollado para procesar texto escrito en idioma inglés, por lo que tuvo un bajo desempeño al momento de codificar eventos a partir de documentos en español. Osorio contactó a Schrodt para explorar la posibilidad de flexibilizar el código de TABARI con el objetivo de mejorar su desempeño para el procesamiento de texto en español. Generosamente, Schrodt accedió a compartir el algoritmo central de TABARI, contribuyendo así a sentar las bases para el desarrollo de EVENTUS ID.

EVENTUS ID y TABARI cuentan con una serie de características técnicas en común tales como el sistema de identificación de palabras y el análisis parcial de frases. Sin embargo, ambos programas tienen particularidades que los diferencian en aspectos fundamentales. Por ejemplo, EVENTUS ID trabaja con un conjunto de algoritmos de codificación de eventos más flexibles que el de TABARI, lo cual se adapta mejor a las complejidades de la lengua española. Además, EVENTUS ID tiene la capacidad de identificar la ubicación geográfica de los eventos con base en la información contenida en el texto fuente. Dicha función no está actualmente incorporada como parte del funcionamiento regular de TABARI. Adicionalmente, dada la irregularidad de la conjugación verbal en español, EVENTUS ID no contiene un sistema automatizado de derivados (*stemming*) como el de TABARI (ver sección 4.3). Si bien la falta de un sistema automatizado de derivados permite reducir errores de codificación de texto en español, también conlleva la necesidad de que el investigador desarrolle diccionarios de codificación detallados y amplios (ver secciones 4.2 y 4.3). En general, la combinación de estas características principales, junto con otras detalladas más adelante, favorecen el desempeño de EVENTUS ID en la codificación y geo-referencia de eventos a partir de texto escrito en español.

Durante el desarrollo de la versión Beta de EVENTUS ID, Leetrau y Schrodt (2013) lanzaron "Global Database on Events, Location and Tone" (GDELT), una enorme colección de más de 200 millones de eventos codificados con TABARI con cobertura en todos los países del mundo desde 1979. Esta base de datos sin precedentes generó inmediatamente el entusiasmo de la comunidad académica acerca del uso de codificación automática para la generación de datos (Ulfedler 2013a,b). Pasado el optimismo inicial, algunos críticos señalaron limitaciones acerca de GDELT y de la codificación computarizada de eventos (Weller y McCubbins 2014; Hanna 2014; Steinert-Threlkeld 2014; Ward et al., 2013). Eventualmente, el debate descansó en torno a una evaluación más balanceada que reconoce el valor de la generación computarizada de eventos a la vez que toma en cuenta sus limitaciones (Moore 2014a,b; Price y Gohdes 2014).

Entre otras limitaciones, los críticos señalan que GDELT sufre un problema de sesgo causado por la cobertura de medios (Beieler 2013).[2] Sin embargo, un aspecto del sesgo de cobertura mediática de GDELT que ha sido poco discutido es el hecho que su codificación está primordialmente basada en fuentes de información escritas en inglés.[3] Al concentrarse exclusivamente en fuentes anglófonas, GDELT ignora una vasta cantidad de notas de prensa con información detallada y oportuna escrita en la lengua nativa de un gran número de países, lo cual reduce la calidad de la información y merma la precisión de la codificación de eventos. Esta situación resulta particularmente preocupante si se considera que aproximadamente el 91.8 por ciento de la población mundial no es angloparlante (Lewis, Simons y Fenning 2013).

En este sentido, EVENTUS ID busca servir como una herramienta para que otros investigadores desarrollen sus propias bases de datos mediante la codificación de eventos a partir de texto escrito en español. En la medida en que este programa contribuya a reducir la brecha de lenguaje y el sesgo de cobertura mediática en el procesamiento computarizado de eventos, las expectativas originales de los desarrolladores de este programa serán excedidas con creces.

1.2 Limitaciones de la codificación automatizada de eventos

La codificación de eventos mediante protocolos computarizados tiene ventajas y desventajas respecto a estrategias tradicionales de codificación manual (Scrhodt y Garner 2012). En términos generales, la codificación automatizada permite reducir sustancialmente los costos de tiempo, trabajo y recursos financieros asociados a proyectos que implican el procesamiento de grandes volúmenes de información. El uso de protocolos computarizados es particularmente conveniente para analizar unidades de texto simples (e.g. párrafos o frases) con construcciones sintácticas

[2] Para una revisión de las características y consecuencias del sesgo de cobertura de medios en investigación sobre conflicto ver Davenport (2009) y Davenport y Ball (2002).

[3] GDELT usa las siguientes fuentes de información: AfricaNews, Agence France Presse, Associated Press, Associated Press Online, Associated Press Worldstream, BBC Monitoring, Christian Science Monitor, Facts on File, Foreign Broadcast Information Service, The New York Times, United Press International, y The Washington Post (ver la sección "Data Sources" en http://gdeltproject.org/about.html#sthash.U65HfSJ3.dpuf).

sencillas y cuando el objetivo de investigación se centra en el contenido literal del texto. Adicionalmente, el procesamiento automatizado de texto permite recodificar fácilmente colecciones enteras de documentos y actualizar la información de manera continua.

En contraste, la codificación humana tiene un mejor desempeño en proyectos donde el interés de investigación se centra en el análisis metafórico o cuando el texto tiene construcciones sintácticas complejas. La codificación humana también suele ser más precisa en proyectos que requieren el análisis integral de documentos enteros o el procesamiento de un volumen reducido de texto. Pese a la confiabilidad de sus resultados, los enormes costos asociados a la codificación humana de grandes volúmenes de documentos suele hacer dichos proyectos poco viables.

Si bien la codificación automatizada de eventos permite procesar grandes volúmenes de texto de manera casi inmediata y a bajo costo, no es plausible esperar que las computadoras codifiquen efectivamente información compleja con la misma precisión que lo harían codificadores humanos. Dado que la codificación automatizada de eventos no es una panacea, resulta fundamental evaluar la precisión del producto generado. Las computadoras no tienen la capacidad de abstracción del cerebro humano. Sin embargo, el procesamiento computarizado de texto puede generar resultados similares a los generados por humanos cuando la tarea de codificación de eventos es simple, literal y discreta (Schrodt y Garner 1994; Best, Carpino y Crescenzi 2013).

Los usuarios de EVENTUS ID, o de cualquier otro programa de codificación de eventos, deben tener en cuenta las limitaciones intrínsecas al procesamiento computarizado de lenguaje natural. Además del desarrollo de diccionarios adecuados, la calidad del producto de codificación depende de manera crucial del grado de complejidad sintáctica del texto. En algunas ocasiones, los textos escritos en español pueden ser demasiados complejos y rebuscados como para generar una codificación precisa. En este tipo de casos, los investigadores podrían optar por la codificación humana si es que cuentan con los recursos necesarios para llevarla a cabo. No obstante, en contextos con recursos limitados, la codificación computarizada puede generar información valiosa. Cualquiera que sea el esquema de codificación (manual o computacional), los investigadores deben señalar las fortalezas y limitaciones del diseño de investigación.

2. PROCESO DE CODIFICACIÓN DE EVENTUS ID

EVENTUS ID utiliza un sistema de reconocimiento de patrones para identificar eventos a partir de texto escrito en español. En esencia, un evento se define como el conjunto de información que describe a alguien haciendo algo a alguien más. Un evento se compone de tres elementos clave:

- **Fuente**: se refiere al actor que inicia la acción o el perpetrador. EVENTUS ID identifica el actor fuente como nombres propios en el texto fuente.

- **Acción**: indica la acción conducida por el actor fuente. El sistema identifica las acciones con base en verbos contenidos en el texto.
- **Objetivo**: se refiere al actor hacia el cual va dirigida la acción iniciada por la fuente.

La estructura de eventos en la secuencia fuente-acción-objetivo corresponde con la estructura gramatical básica de sujeto-verbo-predicado. Esta similitud permite a EVENTUS ID identificar y codificar eventos a partir del contenido de notas de prensa.

Además de indicar quién (fuente) hizo qué (acción) a quién (objetivo), la descripción integral de un evento también debe hacer mención de cuándo y donde ocurrió el evento. Esto constituye dos aspectos adicionales que conforman las características de un evento:

- **Fecha**: se refiere a la fecha cuando ocurrió el episodio.
- **Lugar**: indica la ubicación específica del evento.

Para detectar la fuente, acción y objetivo, EVENTUS ID utiliza diccionarios de actores y verbos. Mientras lee el texto fuente, el programa utiliza la lista de pronombres personales y verbos contenida en los diccionarios como criterios de búsqueda para reconocer actores y acciones. Una vez que estos elementos son detectados, el programa transforma la información textual en formato numérico y lo guarda en una base de datos. Mientras codifica eventos, el programa identifica la fecha de ocurrencia con base en la información provista por el nombre de archivo del texto fuente. Finalmente, el programa recure a un diccionario de locaciones para identificar el lugar donde ocurrió el evento.

2.1 Etapas de codificación

El proceso de codificación de eventos implementado por EVENTUS ID consiste en seis etapas: recopilación de información, formateo del corpus de texto, codificación de eventos, ubicación de eventos, validación del protocolo y generación de archivo de salida. En las primeras dos etapas, los usuarios conforman una colección de documentos relevante cuyo contenido debe ser estructurado de acuerdo a un formato legible por EVENTUS ID. La tercera y cuarta etapas se refieren al proceso de identificación de eventos y lugares. Para garantizar la claridad de exposición, estas dos etapas son discutidas por separado en este manuscrito. Sin embargo, EVENTUS ID ejecuta ambas codificaciones de manera automática y de forma simultánea. La quinta etapa requiere intervención humana para verificar y mejorar la precisión del protocolo de codificación automatizada. La etapa final se refiere a la generación de la base de datos validada. La Figura 1 presenta el diagrama con las diferentes etapas y sus principales tareas. EVENTUS ID codifica eventos de acuerdo con las siguientes seis etapas:

[Insertar aquí Figura 1]

Etapa 1. Recopilación de información. En esta primera etapa los usuarios identifican un conjunto de documentos o notas de prensa relevantes para su investigación. Para la recopilación de información en línea, se recomienda el uso del programa WEB TEXT DOWNLOADER y su procedimiento de extracción y almacenamiento de notas de prensa (Osorio y Reyes 2004b). En los casos en que la información a analizar se encuentre en formato impreso, se recomienda digitalizar la información y procesarla utilizando programas de Reconocimiento Óptico de Caracteres. El objetivo es transformar la información impresa en texto legible para la computadora. El producto de esta etapa es una colección de notas de prensa en archivos individuales. Dada la variedad de técnicas de extracción manual y automatizada de documentos y notas de prensa, este manuscrito no discute a detalle la etapa 1.

Etapa 2. Corpus de texto. En esta etapa, los usuarios utilizan la colección de notas de prensa para construir el corpus de texto que será utilizado por EVENTUS ID como texto fuente. El corpus es un documento de texto que integra y estructura el contenido de todos los reportes de prensa de la colección. Se recomienda el uso del programa auxiliar WEB2EVENTUS y su procedimiento para conformar el corpus de texto de acuerdo al formato requerido por EVENTUS ID (Osorio y Reyes 2004a). El producto de esta etapa es un corpus de texto utilizado como insumo de información para la codificación de eventos. La sección 3 discute a detalle esta etapa.

Etapa 3. Identificación de eventos. En esta etapa, los usuarios utilizan EVENTUS ID para identificar datos de eventos en el corpus de texto. La detección de eventos requiere el desarrollo de diccionarios de actores y verbos. El programa utiliza esas listas de actores y acciones como criterios de búsqueda para identificar los componentes de un evento en el texto analizado: fuente, acción y objetivo. La implementación de esta etapa es realizada por medio de dos algoritmos de identificación de eventos. Esta sección genera una base de datos con los eventos extraídos del corpus de texto. La sección 4 presenta el procedimiento implementado en esta etapa.

Etapa 4. Ubicación de eventos. Esta etapa requiere el desarrollo de diccionarios de locaciones que contengan los nombres de estados y municipios, los cuales son utilizados como criterios de búsqueda. Además, se requiere un diccionario de filtros para evitar ambigüedades en la detección de lugares. El logaritmo de ubicación de eventos utiliza el nombre de las locaciones para inspeccionar el corpus de texto e identificar la locación de los eventos previamente detectados en la tercera etapa. El producto de este procedimiento es una base de datos de eventos geo-referenciados. A pesar que los procedimientos de identificación de lugares y de eventos son analíticamente distintos, EVENTUS ID ejecuta ambos procesos de manera simultánea. La sección 5 discute la ubicación de eventos.

Etapa 5. Validación. En la quinta etapa, los usuarios realizan la codificación manual de una muestra del corpus de texto y la comparan con los datos de eventos generados por EVENTUS ID. Las discrepancias entre la codificación humana y la computarizada sirve para informar la

modificación de diccionarios de actores, verbos y lugares. Se espera que una serie de iteraciones de este procedimiento ayuden a mejorar la precisión del protocolo computarizado y favorezcan la validez de los datos de eventos codificados.

Etapa 6. Producto. El producto final del proceso de codificación automatizada generado por EVENTUS ID consiste en una base de datos validada de eventos geo-referenciados. La etapa de validación y la generación del producto son materia de discusión de la sección 6.

2.2 Requerimientos de sistema y programas auxiliares

El programa EVENTUS ID, el manual de usuario y los archivos de demostración están disponibles para descarga en la siguiente liga:

`http://www.javierosorio.net/#!software/cqbi`

EVENTUS ID funciona en el sistema operativo Windows 7 o superiores. La operación del programa en plataforma Windows permite reducir las barreras tecnológicas para los usuarios en América Latina, donde el uso de Windows es más común que Mac o Unix.

Para ejecutar EVENTUS ID es necesario tener instalado un compilador de perl en versión 5 o superior. Se recomienda el uso de STRAWBERRY PERL, el cual ya contiene las librerías y herramientas para compilar programas en lenguaje Perl. Dicho programa se encuentra disponible en `http://strawberryperl.com/`. Adicionalmente, se requiere contar con un editor de texto para lenguajes de programación. Se recomienda el uso del programa NOTEPAD++, el cual se encuentra disponible en `http://notepad-plus-plus.org/`.

3. CORPUS DE TEXTO

EVENTUS ID está diseñado primordialmente para codificar eventos a partir de información extraída de notas de prensa o breves reportes con características similares. Dicha información sirve para conformar el corpus de texto, el cual es formateado mediante el programa WEB2EVENTUS. Este software auxiliar procesa cada archivo de nota de prensa para extraer su información, partirla por párrafos, adjuntar un contador por cada párrafo, formatear la información y almacenar todo el corpus de texto en un archivo legible para EVENTUS ID.

3.1 Convención de nomenclatura para archivos de insumo

Para procesar documentos de insumo con el programa WEB2EVENTUS es necesario que el nombre de los archivos individuales siga un conjunto de reglas de nomenclatura. Dado que el nombre de archivo constituye el código primario de identificación de cada documento, cumplir

con estos lineamientos es crucial para el funcionamiento adecuado de EVENTUS ID. La nomenclatura de archivos consiste en los siguientes cuatro elementos: fecha, contador, fuente y extensión. El nombre de cada archivo de insumo debe seguir la siguiente estructura: `aaaammddccc_FTE.ext`, donde

- `aaaa` es un número de cuatro dígitos representando el año (e.g. 2009)
- `mm` es un número de dos dígitos representando el mes (e.g. 02 para febrero)
- `dd` es un número de dos dígitos representando el día (e.g. 17)
- `ccc` es un número de tres dígitos contando el número de reportes emitidos por la misma fuente de información en el mismo día. Los tres dígitos permiten que el contador indique hasta 999.
- `FTE` es un acrónimo corto definido por el usuario para indicar el nombre de la fuente de información que emite la nota de prensa.
- `ext` es la extensión que indica el formato del archivo. WEB2EVENTUS es capaz de procesar archivos en formato de *.html* o *.txt*.

No debe haber nombres de archivos duplicados. El contador de la nomenclatura (`ccc`) ayuda a asignar nombres únicos a cada archivo incluso en proyectos que consideren un volumen considerable de notas de prensa. Por conveniencia, el código contador ilustrado en este manuscrito contempla solamente tres dígitos, pero los usuarios pueden asignar el número de dígitos que consideren apropiados para su contador.

Por ejemplo, asuma que el ejército Mexicano, la Secretaría de la Defensa Nacional (SEDENA), emitió tres comunicados de prensa, dos de ellos el día 23 de agosto de 2019 y otro el 17 de octubre de 2010. De acuerdo con los lineamientos de la nomenclatura, los archivos deben tener los siguientes nombres:

```
20090823001_SEDENA.html
20090823002_SEDENA.html
20101017001_SEDENA.html
```

3.2 Formato del corpus de texto

El programa auxiliar WEB2EVENTUS genera un archivo de salida (`corpus.txt`) listo para ser utilizado por EVENTUS ID como texto fuente para la codificación de eventos. Cada fila del archivo de salida contiene información referente a cada uno de los párrafos de las notas de prensa. El contenido del corpus tiene el siguiente formato:

`fecha NombreArchivo_P1_P2 | Texto`, donde:

- `fecha`: indica la fecha de emisión de la nota de prensa, la cual es extraída del nombre del archivo asignado por el usuario (ver lineamientos de nomenclatura en la sección anterior).
- `NombreArchivo`: es el nombre del archivo de cada nota de prensa.
- `P1`: el programa WEB2EVENTUS divide cada nota de prensa en párrafos. El sufijo `P1` es el contador local de párrafos de cada documento. Esto es útil para identificar exactamente de qué párrafo es extraído un evento.
- `P2`: es el contador global de párrafos de todos los reportes que conforman el corpus de texto. Este contador sirve para identificar rápidamente un párrafo en el corpus.
- `|`: el símbolo de barra vertical (`|`) es un marcador que indica el inicio del texto extraído de cada párrafo.
- `Texto`: el contenido de cada párrafo de cada nota de prensa es almacenado en cada línea. La extensión de cada línea depende del número de caracteres de cada párrafo. WEB2EVENTUS no contempla un límite máximo de caracteres para cada línea de texto.

El corpus de texto que utiliza EVENTUS ID consiste en un archivo plano de texto *(.txt*) cuyo contenido está estructurado de la siguiente manera:

```
20130808 20130808001_FTE1 P0 P1 | Lorem ipsum dolor sit amet...
20130808 20130808001_FTE1 P1 P2 | Praesent at sem ac enim ...
20130808 20130808001_FTE1 P2 P3 | Donec sed mattis orci...
20130808 20130808001_FTE2 P0 P4 | Donec velit justo, varius...
20130808 20130808001_FTE2 P1 P5 | Praesent quis felis...
20130808 20130808001_FTE2 P2 P6 | Nunc blandit vitae purus...
20130808 20130808001_FTE2 P3 P7 | Quisque quis lorem sed nunc...
20130921 20130921001_FTE1 P0 P8 | Sed ornare, nisi vitae...
20130921 20130921001_FTE1 P1 P9 | Nulla vel condimentum...
20130921 20130921002_FTE1 P0 P10 | Phasellus porta ipsum eu...
20130921 20130921002_FTE1 P1 P11 | Etiam porttitor vitae odio...
20130921 20130921002_FTE1 P3 P12 | Donec cursus metus vel...
```

En este ejemplo, las primeras tres líneas representan los párrafos extraídos de un reporte de prensa emitido el 8 de octubre de 2013 por la fuente `FTE1`. El contenido de la cuarta a la séptima fila representa los párrafos de un reporte emitido el mismo día por una fuente distinta de información, `FTE2` en este caso. Note que el contador local indica el número de párrafos de cada nota de prensa, mientras que el contador global indica el número consecutivo de párrafos en el corpus de texto. De la octava a la doceava línea, el ejemplo muestra el contenido de dos notas de prensa emitidas por la misma fuente (`FTE1`) el mismo día (21 de septiembre de 2013). Note que la nomenclatura ayuda a asignar identificadores exclusivos por documento (que en este caso son `20130921001` y `20130921002`). Además, los contadores locales y globales permiten identificar cada párrafo en particular.

Además de formatear el corpus de texto con las características requeridas por EVENTUS ID, el programa auxiliar WEB2EVENTUS identifica los acentos diacríticos y enfáticos de las vocales (á, é, í, ó, ú) y los sustituye por la vocal correspondiente sin acento. Si bien las reglas de acentuación son fundamentales para la comunicación oral y escrita en lengua española, existen varias razones que motivan la eliminación de acentos en el corpus de texto de EVENTUS ID. En primer lugar, eliminar acentos facilita la tarea del desarrollo de diccionarios. Dado que hay una gran variedad de formas de escribir caracteres especiales para páginas web en lenguaje de *.html*, los usuarios que quieran usarlos tendrían que realizar la abrumadora tarea de incluir en sus diccionarios palabras con acento utilizando una amplia variedad de códigos especiales (e.g. `arrestó`, `arrest&oacute`, `arrest&#243` o `arrest'f3`). En segundo lugar, desafortunadamente, los errores gramaticales son bastante comunes en notas de prensa escritas en español. En casos así, aun si los usuarios incluyeran palabras con acento en los diccionarios, el programa fracasaría en identificar dichas palabras si el periodista que originalmente escribió la nota omite el uso de acentos. Dado que EVENTUS ID "lee" texto en español sin tildes, los ejemplos de frases o contenido de prensa presentados en este manuscrito carecen intencionalmente de acento y son presentados con texto escrito en fuente `Courier New`.

El proceso de reformateo y limpieza de texto de WEB2EVENTUS también elimina algunos signos de puntuación (e.g. " " : ; ¿ ? ¡ ! - _). Es importante señalar que el programa también crea espacios en blanco a los costados de la coma "," y el punto y seguido ".", quedando de la siguiente forma " . " y " , ". Este tipo de formato es crucial para la localización de eventos de EVENTUS ID detallada en la sección 5.

4. CODIFICACIÓN DE EVENTOS

4.1 Codificación de eventos usando EVENTUS ID

Esta sección explica los pasos básicos para codificar eventos utilizando EVENTUS ID. Los nombres de los archivos mencionados corresponden a aquellos contenidos en el archivo de demostración (EventusID_DEMO.zip) que se encuentra disponible al descargar el programa. EVENTUS ID requiere que los siguientes archivos se encuentren en la misma carpeta de trabajo:

- EVENTUS.pl es el archivo de programa de EVENTUS ID en lenguaje de programación perl.
- actores_DEMO.txt es el diccionario de actores para ser identificados como fuente u objetivo de un evento.
- verbos_DEMO.txt es el diccionario con la lista de acciones a identificar.
- corpus_DEMO.txt es el corpus de texto en formato legible para EVENTUS ID.
- codigos_Eventos_DEMO.txt es el archivo de salida con los eventos identificados por EVENTUS ID y codificados en formato numérico. Adicionalmente, el programa genera un

archivo de salida con la codificación de eventos de manera textual (textos_Eventos_DEMO.txt).

- municipios_DEMO.txt es el diccionario con el nombre de los municipios.
- estados_DEMO.txt es el diccionario que provee la lista de estados para identificar la ubicación de los eventos.
- filtros_DEMO.txt es una lista de nombres para evitar ambigüedades en la detección de lugares.

Con excepción del archivo EVENTUS.pl, que está en formato Perl, todos los demás archivos están en formato de texto plano (*.txt*).

4.1.1 Ejecutando EVENTUS ID paso a paso

Para correr EVENTUS ID de manera manual, los usuarios deben seguir los siguientes pasos:

Paso 1. Abrir la terminal de comando. Ejecutar la interfaz de la terminal de comando de Windows y desplazarse a la carpeta de trabajo que contiene los archivos.[4]

Paso 2. Ejecutar EVENTUS ID. En la interfaz de terminal ingresar los siguientes comandos: `perl EVENTUS.pl` y oprimir la tecla Enter. Esto inicia la operación del programa en el ambiente Perl.

Paso 3. Ingresar la lista de actores fuente. Ingresar el nombre del diccionario de actores usado para identificar la fuente de la acción. Teclear `actores_DEMO.txt` y oprimir Enter.

Paso 4. Ingresar la lista de actores objetivo. Ingresar el nombre del diccionario de actores usado para identificar el objetivo. El programa de demostración utiliza el mismo diccionario de actores para identificar al actor fuente y al actor objetivo. En este caso, escriba nuevamente el nombre del archivo `actores_DEMO.txt` seguido de la tecla Enter. En caso necesario, EVENTUS ID tiene la flexibilidad de utilizar diccionarios de actores diferentes para detectar la fuente y el objetivo de un evento. En este ejemplo, se utiliza el mismo diccionario de actores para identificar ambos componentes de un evento.

Paso 5. Ingrese la lista de acciones. Ingresar el nombre del diccionario de verbos para identificar las acciones realizadas. Escribir `verbos_DEMO.txt` y presionar Enter.

Paso 6. Ingrese el corpus. Indicar el nombre del archivo que contiene el corpus de texto que será utilizado para la identificación de eventos. En la terminal de comando, escribir `corpus_DEMO.txt` y presionar Enter.

[4] Los usuarios no familiarizados con el uso de la terminal de comandos de Windows pueden consultar el siguiente link: http://windows.microsoft.com/en-us/windows/command-prompt-faq" \l "1TC=windows-8

Paso 7. Ingresar la lista de municipios. Indicar el diccionario que contiene el nombre de los municipios a ser utilizados como criterios de búsqueda para detectar la ubicación del evento. Escribir `municipios_DEMO.txt` y oprimir Enter.

Paso 8. Ingresar la lista de estados. Indicar el diccionario que contiene la lista de estados. Ingresar el nombre del archivo `estados_DEMO.txt` seguido de la tecla Enter.

Paso 9. Ingresar la lista de filtros. Escribir el nombre del archivo que contiene la lista de nombres utilizados para evitar ambigüedades en la identificación de locaciones. Ingresar el archivo `filtros_DEMO.txt` y presionar Enter.

Paso 10. Indicar el nombre del archivo de salida. Escribir el nombre de archivo `Eventos_DEMO.txt` y oprimir la tecla Enter. EVENTUS ID utiliza ese nombre para generar dos archivos de salida, uno contiene la codificación numérica de los elementos detectados en el texto fuente (`codigos_Eventos_DEMO.txt`) y el otro contiene la información textual identificada en el corpus (`textos_Eventos_DEMO.txt`).

Paso 11. Seleccionar el algoritmo de codificación. El usuario puede seleccionar dos alternativas para la codificación de eventos. La opción `1` utiliza el algoritmo de codificación general que identifica eventos con la estructura fuente-acción-objetivo. La opción `2` utiliza de manera simultánea el algoritmo general junto con el algoritmo de codificación parcial de eventos con la estructura acción-objetivo. Para detalles sobre ambos algoritmos ver la sección 4.4.

El producto final contiene la información de la fuente, acción, objetivo, fecha y ubicación de cada evento detectado en el corpus de texto. De esta forma EVENTUS ID proporciona información detallada de quién le hizo qué a quién, cuándo y dónde.

4.1.2 Ejecutando EVENTUS ID "rápido y fácil"

Los usuarios pueden ejecutar EVENTUS ID en sólo unos cuantos pasos sin la necesidad de ingresar uno a uno todos los archivos especificados en el procedimiento manual. Para codificar eventos de manera "rápida y fácil" se necesitan dos archivos:

- EVENTUS.pl es el archivo de programa de EVENTUS ID.
- config_DEMO.txt es el archivo de configuración que contiene los nombres de los archivos y entradas necesarias para ejecutar de codificación automática de eventos.

Los nombres de archivos contenidos en config_DEMO.txt corresponden a los documentos señalados en cada uno de los pasos de la ejecución manual del programa. Para ejecutar la codificación automática, dichos archivos deben estar alojados dentro de la misma carpeta. Las entradas de archivo de configuración deben ser listadas en el siguiente orden:

```
actores_DEMO.txt
actores_DEMO.txt
verbos_DEMO.txt
corpus_DEMO.txt
municipios_DEMO.txt
estados_DEMO.txt
filtros_DEMO.txt
Eventos_DEMO.txt
2
```

Para ejecutar EVENTUS ID de forma "rápida y fácil" es necesario implementar los siguientes pasos:

Paso 1. Abrir la terminal de comando. Ejecutar la terminal de comando y desplazarse a la carpeta de trabajo que contiene los archivos.

Paso 2. Ejecutar EVENTUS ID. Ingresar los siguientes comandos: `perl EVENTUS.pl config_DEMO.txt` seguido de la tecla Enter. El primer elemento de este comando invoca el ambiente Perl, el segundo activa el programa EVENTUS ID y el tercero provee el nombre de todos los demás archivos requeridos.

El último elemento del archivo de configuración corresponde a la opción `3` del Paso 12, la cual genera tanto el archivo de codificación numérica como la codificación textual. Como se indicó anteriormente, lo usuarios pueden ingresar los números 1, 2 y 3 para seleccionar el tipo de archivo de salida deseado.

El resto de esta sección discute las características de los diccionarios de actores y verbos, así como los diferentes algoritmos de codificación de eventos utilizados por EVENTUS ID. Más adelante, la sección 5 discute el procedimiento de identificación de lugares.

4.2 Diccionario de actores

EVENTUS ID utiliza diccionarios de actores para identificar la fuente y el objetivo de un evento. El programa tiene la flexibilidad de detectar cada uno de estos elementos utilizando un diccionario diferente para el actor fuente y otro para el actor objetivo. Sin embargo, para facilitar la explicación, esta sección asume el uso de un solo diccionario de actores para codificar tanto la fuente como el objetivo. El diccionario de actores consiste en una lista de pronombres personales que hacen referencia al actor que inicia la acción o al destinatario de las mismas. Estos pronombres personales sirven de criterios de búsqueda para detectar los actores en el corpus de texto. Cada elemento de la lista de actores está asociado a un código numérico que indica la categoría a la que pertenece cada actor. El código debe ir entre corchetes.

El diccionario de actores es un archivo de texto plano (*.txt*). Los nombres de actores compuestos por varias palabras van separados por un guión bajo "_" para ayudar al programa a identificar palabras en el texto. El guión bajo también debe ser el último elemento al final de cada pronombre personal. EVENTUS ID no requiere que las palabras del corpus de texto estén separadas por guiones bajos. El programa "lee" el corpus que contiene palabras separadas simplemente por espacios en blanco como en cualquier otro texto, pero utiliza diccionarios de actores con palabras concatenadas por un guión bajo para identificar patrones en el texto fuente. Una vez que el nombre de un actor ha sido identificado en el corpus, el código correspondiente es almacenado. La siguiente lista presenta un ejemplo del diccionario de actores:

```
tropas_del_ejercito_ [202051]
oficial_de_policia_ [204021]
miembros_de_una_organizacion_criminal_ [601060]
cocaina_ [801022]
AK_47_ [901013]
```

4.3 Diccionario de verbos

El diccionario de verbos consiste en una lista de conjugaciones verbales que hacen referencia a las acciones realizadas por el actor fuente o dirigidas hacia en actor objetivo. Tal como el diccionario de actores, la lista de verbos debe estar contenida en un archivo de texto plano (*.txt*). Cada elemento de la lista de verbos es seguido por un código numérico referente a la categoría de acción asignada por el usuario. Los códigos numéricos deben ir entre corchetes. En algunos casos, las acciones pueden ser efectivamente identificadas con un simple verbo. En otros casos, las complejidades del lenguaje natural obligan al uso de frases más complejas para identificar adecuadamente las acciones de interés. Las acciones compuestas por múltiples palabras deben ir separadas por un guion bajo "_", el cual también es requerido al final de cada frase verbal. A continuación se presenta un ejemplo del diccionario de verbos:

```
ataca_ [88101]
atacar_ [88101]
atacados_ [88101]
- fueron_ * [99101]
arresto_ [88104]
- bajo_ * [88104]
combate_ [88101]
- fortalecer_el_*contra_ [- - -]
```

Con el objetivo de mejorar la precisión del protocolo de codificación, los usuarios deben considerar el uso de una gran variedad de conjugaciones verbales para referirse al mismo tipo de acción. Como lo muestra el ejemplo anterior, la sintaxis de los verbos "`ataca`", "`atacar`" y

"atacados" corresponde a diferentes conjugaciones del mismo verbo, por lo tanto tienen el mismo código numérico [88101] para referir a la misma acción.

Algunos otros verbos van seguidos de un conjunto de palabras relacionadas que ayudan a precisar el significado de la acción. En estos casos, el símbolo * sirve como un comodín indicando el lugar de la frase donde debe aparecer el verbo mencionado con anterioridad. En el ejemplo anterior, el programa inserta el verbo "atacados" en la frase fueron_* y busca la conjugación verbal "fueron atacados" en el corpus de texto. Note que el código del verbo "atacados" es ligeramente diferente al de "fueron atacados. En ambos casos la raíz del código es 101, pero el primero inicia con el prefijo 88 y el segundo con 99. El prefijo 99 es utilizado como un sistema para evitar ambigüedades en la conjugación de verbos en voz pasiva. Esta referencia sirve como clave durante el proceso de validación y recodificación, así como para el análisis estadístico posterior a la codificación de eventos. Por ejemplo, en la frase "un oficial de policía fue atacado por miembros de una organización criminal" el programa identifica la secuencia fuente-acción-objetivo como 204021 → 99101 → 601060. El cual hace referencia a los elementos "policía" → "fue atacado" → "organización criminal". Sin embargo, data la secuencia de eventos fuente-acción-objetivo, esta codificación podría ser erróneamente interpretada como un episodio en el que la policía atacó a una organización criminal, lo cual invertiría la direccionalidad del evento. Para evitar este tipo de confusiones, el prefijo 99 ayuda al usuario a identificar construcciones verbales escritas en voz pasiva e invertir el orden de la fuente y objetivo del evento en la etapa de recodificación para que el código quede de la siguiente forma 601060 → 99101 → 204021 y el evento se pueda interpretar como "organización criminal" → "atacó" → "policía" (ver sección 6.2).

Como se mencionó anteriormente, EVENTUS ID utiliza un esquema de codificación basado en el reconocimiento de patrones similar al de TABARI. Sin embargo, la principal diferencia entre estos dos programas radica en la forma en que procesan las conjunciones verbales. TABARI está diseñado para procesar texto escrito en inglés, cuyas reglas gramaticales permiten una construcción simple y general de frases con la estructura sujeto, verbo y objetivo. Por ejemplo, la mayoría de los verbos en inglés pueden ser fácilmente conjugados al añadir las terminaciones "s", "ed" o "ing" al final del verbo en su forma infinitiva. Otra característica del idioma inglés es que el género y el número de un pronombre generalmente no afectan la conjugación de los verbos. Esto significa que ese posible combinar las personas "you", "he", "she", "we" y "they" con cualquier conjugación verbal de manera indistinta. Con base en estas simples reglas gramaticales, TABARI utiliza un sistema automatizado de derivados (*stemming*). Utilizando el verbo arrestar (*to arrest*) como ejemplo, la Tabla 1 muestra como el idioma inglés permite conjugar fácilmente el verbo arrestar en diferentes tiempos verbales para distinto género y número simplemente añadiendo los sufijos "s", "ed" o "ing" al final del verbo en infinitivo.

[Insertar aquí Tabla 1]

Aunque el sistema de derivados de TABARI es conveniente para la codificación de texto escrito en inglés, este tipo de función genera errores de codificación en el procesamiento de texto escrito en idioma español. Utilizar un sistema de derivados basado en inglés no es apropiado para codificar eventos en español dado que los verbos en este idioma no terminan con "s", "ed" o "ing". Para exponerlo de la manera más simple posible, idiomas diferentes requieren procesos de codificación diferentes. La Tabla 1 muestra las diferentes conjugaciones del verbo arrestar a través de diferentes tiempos verbales, género y número de personas.

Dadas las complejidades de la conjugación verbal en español, EVENTUS ID no incluye un sistema automatizado de derivados. La ausencia de una función que le permita al programa identificar derivados disminuye la propensión de error en la codificación de eventos. Sin embargo, esto implica que el usuario debe desarrollar diccionarios de verbos lo más amplios y detallados posible para incluir una gran variedad de conjugaciones verbales para diferente género y número de personas. Es importante recordar que el protocolo de codificación utiliza la información de los diccionarios como criterios de búsqueda, por lo tanto el usuario debe proveer dichos criterios con el mayor detalle posible para lograr una codificación precisa.

El desarrollo de diccionarios de verbos y actores requiere un proceso gradual de aprendizaje, conocimiento acumulado, lectura cuidadosa y retroalimentación del proceso de validación. Este proceso interactivo de codificación, verificación y recodificación permite afinar los diccionarios mediante la inclusión de nuevos actores y verbos o la modificación de los ya existentes. Sin embargo, es importante señalar que no es posible desarrollar diccionarios perfectos que permitan codificar con exactitud absolutamente todos los eventos y logar un 100 por ciento de precisión. Dadas las complejidades del lenguaje natural, es necesario reconocer que el protocolo de codificación genere cierto margen de error. A pesar de esta limitación, es posible estimar el rango de error en el procesamiento de eventos e incorporarlo como medida de incertidumbre en el análisis de datos. Como lo indican Grimmer y Stewart (2013), todos los modelos cuantitativos de lenguaje natural tienen fallas; sin embargo, algunos son útiles.

En ocasiones, los usuarios pueden estar interesados en codificar palabras referentes a actores o verbos que tienen significados ambiguos dependiendo del contexto específico de la frase en que se encuentren. Este tipo de ambigüedades pueden generar errores de codificación. Para estos casos, EVENTUS ID permite la posibilidad de ignorar ciertos elementos de los diccionarios de actores y verbos que pudieran generar la codificación equivocada de eventos. Por ejemplo, usuarios interesados en codificar confortamientos armados entre autoridades gubernamentales y organizaciones criminales pueden incluir en el diccionario el verbo "`combate`" con el código `[88101]`. Sin embargo, los reportes de prensa suelen terminar con elementos retóricos como la siguiente frase: "`Estas acciones son muestra del`

esfuerzo del gobierno para fortalecer el combate contra el crimen organizado". En este contexto, el verbo "combate" tiene un sentido figurado que hace referencia a las acciones gubernamentales en contra del crimen organizado, pero no se refiere a la ocurrencia de un enfrentamiento armado entre agentes de seguridad del estado y grupos criminales. Para evitar la codificación errónea de este tipo de frases, EVENTUS ID tiene un mecanismo de resolver ambigüedades. Los usuarios deben incluir el código nulo [---] para indicarle al programa las palabras o frases que deben ser ignoradas en el proceso de codificación. En el ejemplo anterior, los usuarios pueden incluir en el diccionario de verbos la frase "fortalecer el combate contra" [---] para ignorar este fragmento y evitar un error de identificación de eventos. Este código nulo es aplicable para elementos de los diccionarios de actores y de verbos.

4.4 Algoritmos de codificación de eventos

EVENTUS ID utiliza dos algoritmos para codificar eventos: el *algoritmo general* codifica eventos que contienen la secuencia completa de fuente-acción-objetivo y el *algoritmo parcial* codifica eventos con la secuencia incompleta de acción-verbo. Ambos algoritmos utilizan la técnica de análisis parcial de frases (*sparse parsing*) desarrollada por Schrodt en KEDS y posteriormente en TABARI. El método de análisis parcial de frases utiliza los diccionarios de actores y verbos como criterios de búsqueda para identificar solamente las partes relevantes del texto que corresponden a un evento, mientras que el resto del texto es ignorado para propósitos de codificación.

Con base en la nomenclatura de archivos descrita en la sección 3.1, ambos algoritmos toman la información del corpus de texto e identifican primero la fecha del evento (fecha) y después el nombre del documento y su párrafo correspondiente (NombreArchivo_P1_P2). Posteriormente, cada algoritmo utiliza su propio esquema de codificación para reconocer los eventos contenidos en el texto fuente. La Tabla 2 muestra los pasos implementados por cada algoritmo para la codificación de eventos. Finalmente, EVENTUS ID almacena los eventos codificados en una base de datos contenida en un archivo de texto plano (*.txt*). Cada línea del archivo de salida contiene un conjunto de elementos que corresponden al evento codificado, separando sus componentes con un espacio tabular. Como resultado, los algoritmos general y parcial generan productos con las siguientes características:

```
fecha → NombreArchivo_P1_P2 → actor1 → verbo → actor2
fecha → NombreArchivo_P1_P2 → → verbo → actor2
```

[Insertar aquí Tabla 2]

4.4.1 Algoritmo general

El algoritmo de secuencia general de EVENTUS ID identifica eventos que presentan la secuencia completa de fuente-acción-objetivo. Para codificar eventos con este algoritmo, es necesario que los tres elementos se encuentren en el orden preciso en el texto fuente. Como ejemplo, considere la siguiente frase: "Tropas del ejercito arrestaron a miembro de una organizacion criminal". En este ejemplo, los tres elementos están presentes en el orden requerido. Por lo tanto, el programa identifica las "tropas del ejercito" como actor fuente, el verbo "arrestaron" como la acción realizada y "miembro de una organizacion criminal" como actor objetivo. El producto de la codificación almacena en la base de datos el siguiente código numérico: 202051 → 88104 → 601060.

Dado que el análisis parcial de frases sólo se enfoca en los elementos relevantes del texto provistos por los diccionarios de actores y verbos, el texto a procesar puede ser más extenso sin que eso afecte el resultado de la codificación. Por ejemplo, considere el siguiente párrafo:

En un comunicado de prensa emitido el dia de ayer, el gobierno mexicano informo que tropas destacamentadas en el municipio de San Luis Rio Colorado, Son. decomisaron paquetes de mariguana con un peso total de dos toneladas y 250 gramos, mientras patrullaban caminos rurales del area.

Pese a la longitud del párrafo, el análisis parcial de frases le permite a EVENTUS ID reconocer los componentes claves del evento y codificar a las "tropas" como la fuente, "decomisaron" como el verbo y "mariguana" como el objetivo.

Como se mencionó anteriormente, dadas las características del estilo periodístico latinoamericano, es bastante común encontrar el uso de voz pasiva y presente indicativo en la redacción de reportes de prensa. La voz pasiva incrementa la complejidad gramatical de las frases al invertir el orden del sujeto y el predicado. Por ejemplo, considere la siguiente frase: "Un miembro de una organizacion criminal fue arrestado por tropas del ejercito". Los tres componentes de un evento están presentes en esta frase. Sin embargo, la fuente y el objetivo se encuentran en orden inverso. De acuerdo al algoritmo general, EVENTUS ID codificaría "miembro de una organizacion" como el primer actor, el verbo "arrestado" como la acción y "tropas del ejercito" como segundo actor y generaría el código: 601060 → 99104 → 202051.

Sin embargo, el orden de esta codificación podría ser interpretada como "un miembro de una organización criminal arrestó a tropas del ejército", lo cual no corresponde a la idea presentada en el texto. Para evitar este tipo de confusiones, el código del verbo incluye el prefijo 99, el cual ayuda a identificar verbos en voz pasiva. Este prefijo puede ser utilizado como clave para corregir la direccionalidad del evento en el proceso de validación y recodificación (ver

sección 6.2). De esta forma, el desarrollo de diccionarios, los algoritmos de codificación y el esquema de recodificación trabajan en conjunto para procesar construcciones gramaticales complejas y reducir los errores de codificación en la base de datos. Utilizando una simple regla de recodificación para invertir el orden de los actores con verbos que indiquen voz pasiva, el resultado de la recodificación puede ser el siguiente: `202051 → 88104 → 601060`.

Como lo indica la Tabla 2, el algoritmo de secuencia general inicia la búsqueda por el actor fuente; una vez que es detectado, el algoritmo cambia a la búsqueda de verbos y finalmente busca el actor objetivo. Sin embargo, en algunas ocasiones los reportes mencionan una serie de elementos que no son seguidos de verbos. Esto es común en comunicados de prensa del gobierno que listan artículos decomisados en operativos policiacos. Considere el siguiente ejemplo:

```
Tropas del ejercito arrestaron a miembro de un grupo criminal.
Las tropas decomisaron 6 kilogramos de cocaina y los siguientes
articulos:
- 372 paquetes de fosfato de clindamycina
- dos AK-47
- un rifle de asalto R-15
- municiones de varios calibres
```

Aplicando el algoritmo general, EVENTUS ID identificaría el primer evento con los siguientes componentes: "`tropas del ejercito`" como la fuente, el verbo "arrestaron" como la acción y "`un miembro de un grupo criminal`" como el actor objetivo. En el segundo evento la fuente es "las tropas", la acción es "`decomisaron`" y "`cocaina`" es el objetivo. En la tercera línea, el algoritmo inicia buscando el primer actor e identifica "`fosfato de clindamycina`" como la fuente. Dado que ya no hay más verbos, el algoritmo termina la búsqueda en esa línea y pasa a la siguiente e inicia nuevamente con la búsqueda del actor. De esta forma, el algoritmo general detectaría "`AK-47`", "`rifle de asalto R-15`" y "`municiones`" como elementos independientes en cada línea. El resultado de la codificación quedaría de la siguiente forma:

```
202051 → 88104 → 601060
202051 → 88202 → 801022
604021 → →
901013 → →
901014 → →
901021 → →
```

Finalmente, se puede recodificar fácilmente la lista de elementos decomisados como el actor objetivo e imputar la información del actor fuente y el verbo mencionado al inicio del párrafo.

4.4.2 Algoritmo parcial

El algoritmo de secuencia parcial de EVENTUS ID sirve para procesar estructuras gramaticales más complejas, como aquellas contenidas en frases escritas en presente indicativo. El presente indicativo se construye removiendo la parte infinitiva del verbo (e.g. eliminar la terminación "ar" en el verbo "arrestar") y reemplazándola con la terminación que indica la persona que ejecuta la acción. De esta forma, el presente indicativo omite el sujeto de la acción e inicia la frase con el verbo, seguido del predicado. Por ejemplo, la frase "arrestan a un criminal" está escrita en presente indicativo y hace referencia a que un criminal fue arrestado, pero no revela quién realizó el arresto. La complejidad del presente indicativo radica en que la conjugación del verbo ya incluye información acerca de la persona.

Adicionalmente, el presente indicativo es frecuentemente utilizado para referirse a eventos que ocurrieron en el *pasado* (presente histórico). De tal forma, la frase "arrestan a un criminal" literalmente se refiere a una acción que ocurre en tiempo presente, pero en el estilo periodístico dicha frase también hace referencia figurativa a un evento ocurrido en el pasado. El uso de presente indicativo es sumamente común en la estructura gramatical utilizada en la narrativa periodística latinoamericana (Martínez, Miguel y Vázquez 2004, Alcoba Rueda 1983, Nadal Palazón 2009). De acuerdo a Guízar García (2004), aproximadamente el 73 por ciento de las notas de prensa en México usan el presente indicativo en sus encabezados.

El algoritmo parcial ayuda a codificar frases en las que el verbo es conjugado en presente indicativo. Por ejemplo, considere la siguiente frase: "`Arrestan a un criminal`". En esta frase, la conjugación en presente indicativo omite el sujeto de la acción. Ante la ausencia del primer actor, EVENTUS ID recurre al algoritmo de secuencia parcial para identificar primero el verbo "`arrestan`" como la acción realizada y "`a un criminal`" como objetivo de la misma. Como resultado, el programa generaría el siguiente código: `· → 88104 → 601060`, donde el signo · representa un espacio en blanco. Como se discute en la sección 6, los usuarios pueden desarrollar protocolos de recodificación de eventos para llenar los espacios vacíos generados a partir frases en presente indicativo.

5. CODIFICACIÓN DE LUGARES

Las características de EVENTUS ID anteriormente mencionadas permiten extraer información del corpus de texto referente a la fecha del evento, el actor fuente que inicia una acción, el tipo de acción realizada y el objetivo de la misma. Sin embargo, para tener un recuento completo del suceso, es indispensable saber dónde ocurre el evento. EVENTUS ID cuenta con una función plenamente integrada que le permite reconocer en el texto fuente el lugar donde ocurrió el evento considerando dos niveles de desagregación sub-nacional (estados y municipios). Si bien esta

sección discute por separado el proceso de geo-referenciación de eventos, EVENTUS ID implementa esta función de manera automática durante el proceso de identificación de eventos.

En términos generales, el protocolo de identificación de lugares funciona de la siguiente manera. El programa utiliza los diccionarios de estados y municipios como criterios de búsqueda para reconocer el lugar de ocurrencia de un evento en el corpus de texto. Además, debido a la complejidad en la identificación de lugares, el protocolo de geo-referenciación incluye un diccionario de filtros que permite resolver ambigüedades y evitar que algunas palabras o frases sean erróneamente identificadas como localidades.

5.1 Diccionarios de lugares

El protocolo de geo-referenciación de eventos de EVENTUS ID requiere el desarrollo de dos tipos distintos de diccionarios, uno con el listado de estados y otro con el listado de municipios. Ambos diccionarios deben estar contenidos en archivos con formato de texto plano (*.txt*). El programa utiliza estos diccionarios como categorías de búsqueda para el reconocimiento de patrones en el corpus de texto. Cada línea de los diccionarios de lugares debe iniciar con el código numérico del lugar, seguido por el nombre textual de cada locación. Estos dos elementos deben ir separados por un espacio tabular (→).

La siguiente lista presenta un ejemplo del diccionario de estados:

```
1 → Aguascalientes
2 → Baja California
3 → Baja California Sur
4 → Campeche
5 → Coahuila
```

Este es un ejemplo del diccionario de municipios:

```
1002 → Asientos
2004 → Tijuana
3001 → Comondu
4010 → Calakmul
5025 → Piedras Negras
```

El diccionario de filtros sirve para evitar que el algoritmo de localización de eventos genere falsos positivos. En la investigación sobre violencia del crimen organizado en México que motivó el desarrollo de EVENTUS ID, Osorio (2013) señala que algunos problemas de ambigüedad en la identificación de lugares surgen del hecho que varios grupos criminales tienen el nombre de las ciudades o estados donde operan. Tal es el caso de “El Cártel de Tijuana”, “El Cartel de Juárez” o el “Cártel de Sinaloa”, entre otros. EVENTUS ID utiliza el diccionario de

filtros para reducir el riesgo que el algoritmo codifique erróneamente el nombre de organizaciones criminales como lugares de ocurrencia de eventos. De manera similar, el diccionario de filtros ayuda a evitar errores de localización derivados de periódicos que llevan el nombre de la ciudad donde circulan. Tal es el caso de periódicos como "El Diario de Juárez" o "El Sol de Puebla," entre otros.

Otra posible causa de ambigüedad en la identificación de lugares proviene del encabezado y pie de página de los comunicados de prensa de agencias gubernamentales, ya que estos generalmente mencionan la dirección postal de la agencia que los emite. Dicha información puede generar confusión en la codificación de lugares al relacionar erróneamente la ocurrencia del evento con la dirección de la agencia emisora. La inclusión de dichas direcciones en el diccionario de filtros permite eliminar estos falsos positivos. En otros casos, los errores de localización surgen de notas de prensas que narran la intercepción de algún vehículo por parte de las autoridades gubernamentales. Por ejemplo, este tipo de reportes suelen mencionar que un vehículo con placas del estado "X" fue detenido por las autoridades en la carretera del estado "Y," mientras transitaba con dirección al estado "Z". El diccionario de filtros ayuda a minimizar el riesgo que este tipo de situaciones generen errores de localización.

Al igual que los demás diccionarios, la lista de filtros debe estar en un archivo con formato de texto plano (*.txt*). El primer elemento de cada línea debe ser el código numérico `0`, seguido por un espacio tabular (→) y posteriormente el texto correspondiente a la categoría de exclusión. La asignación del código `0` le indica a EVENTUS ID la instrucción de ignorar dicho elemento en caso de ser encontrado en el corpus de texto.

La siguiente lista muestra un ejemplo del diccionario de filtros:

```
0 → Cartel de Sinaloa
0 → Cartel de Juarez
0 → Cartel de Tijuana
0 → Zona Militar La Paz BCS
0 → Operativo Conjunto Michoacan
0 → Operacion Conjunta Nuevo Leon
```

5.2 Ubicación de lugares usando EVENTUS ID

La Tabla 3 presenta el protocolo de localización de eventos de EVENTUS ID. En general, el algoritmo de localización utiliza como insumo la base de datos de eventos codificados e identifica el nombre del párrafo del que se extrajo cada evento. Después, el programa lee todo el corpus de texto para identificar el párrafo que contiene el evento identificado en la base de datos. Una vez que el párrafo es detectado, el algoritmo utiliza la información de los diccionarios de estados y municipios para identificar el lugar de ocurrencia del evento en el párrafo de origen. Si la locación es identificada en el párrafo, el protocolo recurre al diccionario de filtros para

verificar si la localidad debe ser registrada o descartada. Si la ubicación no es descartada, el algoritmo almacena en la base de datos el código del estado y/o municipio junto al código del evento previamente identificado. Si la locación no es identificada en el párrafo originario, el algoritmo expande la búsqueda de lugares al resto del documento iniciando por el primer párrafo del mismo. Si el programa identifica la ubicación del evento en algún otro párrafo del documento, el protocolo verifica si dicha ubicación debe ser filtrada o no. En caso que la localidad pase el filtro, el algoritmo imputa el estado y/o municipio en la línea del evento analizado y guarda sus códigos junto a los del evento previamente identificado. El protocolo de codificación es capaz de registrar el nombre de múltiples estados y municipios cuando estos son mencionados en el párrafo o documento correspondiente.

[Insertar aquí Tabla 3]

Para ilustrar la identificación de lugares de EVENTUS ID considere el siguiente párrafo:

`En un comunicado de prensa emitido el dia de ayer, el gobierno mexicano informo que tropas destacamentadas en el municipio de San Luis Rio Colorado, Son. decomisaron paquetes de mariguana con un peso total de dos toneladas y 250 gramos, mientras patrullaban caminos rurales del area.`

Asumiendo los diccionarios de actores, verbos y lugares adecuados, EVENTUS ID sería capaz de reconocer los componentes clave del evento tales como las "`tropas`" (fuente), "`decomisaron`" (acción) y "`paquetes de mariguana`" (objetivo). Adicionalmente, el algoritmo de ubicación de lugares identificaría el municipio de "`San Luis Rio Colorado`" en el estado de "`Son.`" (Sonora) como el lugar donde ocurrió el evento.

5.3 Generación de datos geo-referenciados

El producto del proceso de codificación de eventos y del protocolo de geo-referenciación es una base de datos en formato de texto plano (*.txt*) que indica la fecha de ocurrencia, la fuente del evento, la acción realizada y el objetivo hacia el cual va dirigida, así como la ubicación del estado y municipio donde ocurrió dicho evento. EVENTUS ID genera automáticamente dos versiones del archivo de salida. Un archivo contiene los códigos numéricos de los eventos y localidades detectadas (codigos_ArchivoSalida.txt) y el otro archivo contiene la información en formato textual (textos_ArchivoSalida.txt). De esta forma EVENTUS ID genera información detallada acerca de quién le hizo qué a quién, cuándo y dónde a partir del texto fuente.

A continuación se presenta un ejemplo del formato genérico del archivo de salida de EVENTUS ID. Los elementos `estado1` y `mun1` pueden ser códigos numéricos o anotaciones textuales dependiendo del tipo de archivo de salida.

```
fecha1 NombreArchivo_P1_P1 actor1 verbo actor2 estado1 mun1
fecha1 NombreArchivo_P2_P2 actor1 verbo actor2 estado1 mun1
fecha1 NombreArchivo_P3_P3 actor1 verbo actor2 estado1 mun1
fecha2 NombreArchivo_P1_P4 actor1 verbo actor2 estado1 mun1
fecha2 NombreArchivo_P2_P5 actor1 verbo actor2 estado1 mun1
```

6. VALIDACIÓN DE EVENTOS Y RECODIFICACIÓN

La generación de medidas válidas es una preocupación central en las ciencias sociales, tanto para la investigación que utiliza métodos cuantitativos como cualitativos (King, Keohane and Verba, 1994; Adcock and Collier, 2001; Collier and Brady, 2004; Bollen, 1989; Goertz, 2005). La validez de una medición se cumple cuando las métricas capturan de manera fehaciente las ideas contenidas en los conceptos que intentan medir. Como lo indican Adcock y Collier (2001), una medición es válida en función de que sus métricas correspondan a un conjunto de indicadores que pueden ser interpretados en términos de la definición usada para representar un concepto.

EVENTUS ID es un instrumento de generación de métricas que pueden ser utilizadas para la elaboración de indicadores. Desde una perspectiva elemental, la base de datos generada de manera computacional es válida en la medida que los diccionarios y el protocolo de codificación identifiquen de manera precisa los eventos de interés en el corpus de texto. Un criterio de validez más amplio debe considerar también la congruencia entre la base de datos computarizada y las construcciones conceptuales definidas por el investigador, así como la sistematización y desarrollo de indicadores. Los siguientes apartados presentan recomendaciones para evaluar el grado de validez de la codificación computarizada de eventos desde una perspectiva elemental. Sin embargo, dado que cada proyecto de codificación tiene sus propias motivaciones teóricas, esta sección no profundiza en la evaluación de validez en el sentido amplio.

6.1 Proceso de validación

¡Valida, valida, valida! Esta es la principal recomendación que ofrecen Grimmer y Stewart (2013) para el procesamiento computarizado de lenguaje natural. En términos generales, el objetivo del proceso de validación es reducir el riesgo de errores tipo I y II. Los usuarios pueden reducir el riesgo de identificar falsos negativos (error tipo II) al desarrollar listas detalladas y comprehensivas de actores, verbos y lugares; capaces de identificar los comportamientos de interés en el texto fuente. Además, los usuarios pueden desarrollar un conjunto de excepciones a

dichos diccionarios para reducir el riesgo de falsos positivos (error tipo I) que pudieran generar la codificación errónea de un evento cuando éste no tuvo lugar.

Los usuarios pueden seguir el siguiente procedimiento básico para evaluar la precisión y validez de la codificación automatizada de eventos:

Paso 1. Generación de codificación humana. Primero seleccione una muestra aleatoria de documentos o párrafos del corpus de texto. Posteriormente, recurra a un equipo de codificadores humanos para identificar los eventos contenidos en el corpus de muestra. Los codificadores deben seguir el procedimiento de los algoritmos de identificación de eventos y lugares de EVENTUS ID (ver secciones 4.4.1, 4.4.2 y 5.2). Dada la complejidad del lenguaje natural, la base de datos generada por humanos es considerada como el estándar de codificación para evaluar la validez de los datos generados de manera computacional.

Paso 2. Comparar la codificación humana y la automatizada. Recurra a EVENTUS ID para generar una base de datos a partir del corpus de muestra. Posteriormente, compare cada evento identificado por el equipo de codificadores humanos con la codificación computarizada. Las discrepancias entre la base de datos humana y la computacional deben ser utilizadas para modificar y mejorar el protocolo de codificación automatizado.

Paso 3. Ajustar el protocolo de codificación. EVENTUS ID ofrece amplia flexibilidad para que los usuarios implementen el sistema de codificación que mejor se ajuste a sus necesidades. Con base en las discrepancias identificadas en el paso anterior, los usuarios pueden considerar las siguientes opciones para mejorar la precisión del protocolo de codificación automatizada:

- **Mejorar los diccionarios de actores y verbos.** La forma más común de aumentar la precisión del protocolo de codificación es ajustando los diccionarios de actores y verbos. Esto generalmente implica la inclusión de pronombres y verbos que no fueron considerados en versiones anteriores de los diccionarios. Los usuarios también pueden recurrir al código nulo [- - -] para evitar ambigüedades en casos específicos.
- **Considere el uso de diferentes diccionarios de actores.** EVENTUS ID ofrece la posibilidad de utilizar distintos diccionarios de actores para identificar la fuente y el objetivo de un evento. En caso que los usuarios decidan usar diccionarios diferenciados, es importante que tomen en cuenta la secuencia de identificación de actores implementada por los logaritmos general y parcial del programa, así como la complejidad sintáctica del corpus de texto.
- **Seleccione el algoritmo de codificación de eventos apropiado.** El programa permite a los usuarios codificar eventos utilizando solamente el algoritmo de secuencia general (fuente-acción-objetivo) o incorporar el esquema de codificación del algoritmo parcial (acción-objetivo). Se recomienda realizar codificaciones de prueba para determinar si el uso de uno o ambos algoritmos es más conveniente para el proyecto en curso.

- **Mejorar los diccionarios de lugares y filtros.** La identificación de lugares de manera precisa puede ser una tarea difícil de realizar. Es posible que los usuarios necesiten aumentar y refinar los diccionarios de estados y municipios para mejorar la capacidad de identificación de lugares de ocurrencia. Quizá de manera más importante, es posible que los investigadores tengan que desarrollar un conjunto confiable de filtros para evitar la codificación errónea de lugares.

Como lo muestra la Figura 1, el proceso de validación consiste en un ciclo de retroalimentación entre los protocolos de codificación, validación, ajuste y recodificación. Después de generar el estándar de codificación humana y el primer producto de codificación computarizada, los usuarios evalúan las discrepancias entre ambas codificaciones. Dicha información sirve para mejorar el protocolo de codificación computacional, el cual nuevamente es evaluado a la luz de la codificación humana. La repetición de este protocolo de evaluación busca incrementar la convergencia entre las bases de datos generadas con codificación humana y automatizada. Sin embargo, es importante que los usuarios tengan en mente que es sumamente difícil (si no imposible) generar un sistema de codificación que sea 100 por ciento perfecto.

6.2 Recodificación de eventos

Después de validar los eventos codificados, algunos usuarios posiblemente consideren realizar codificaciones adicionales para mejorar la precisión de la base de datos. EVENTUS ID no incluye una función para recodificar eventos automáticamente, pero los usuarios pueden manipular fácilmente los datos utilizando programas de análisis estadístico como STATA y R. Los archivos de demostración de EVENTUS ID incluyen un ejemplo de recodificación y agregación de eventos en un dofile de STATA.

La recodificación de eventos puede ser particularmente necesaria para modificar el orden de eventos extraídos de frases escritas con voz pasiva. Como lo indican las secciones 4.3 y 4.4, la voz pasiva invierte la posición sintáctica del sujeto y el predicado, lo cual puede dar paso a la interpretación errónea del evento codificado cuando es analizado fuera de contexto.

Por ejemplo, considere la siguiente frase en voz pasiva: "Un miembro del crimen organizado fue arrestado por tropas militares." La construcción sintáctica de esta frase presenta primero el predicado, posteriormente el verbo y finalmente el sustantivo. Dado el orden de los elementos en la estructura de la frase, EVENTUS ID codificaría la información de manera textual como: "miembro del crimen organizado" → "fue arrestado" → "tropas militares", y de manera numérica como: 601060 → 99104 → 202051. Al enfocarse exclusivamente en los códigos numéricos (tal como lo hace el análisis estadístico), esta codificación no sería capaz de reflejar de manera precisa la direccionalidad del evento. De hecho, la lectura descontextualizada del código numérico podría

sugerir erróneamente que “un miembro del crimen organizado” (601060) “arrestó” (99104) a “tropas militares” (202051), lo cual no corresponde con la idea presentada en la frase.

Para evitar este tipo de errores de codificación, los usuarios pueden desarrollar protocolos de recodificación para revertir la estructura predicado-verbo-sujeto de la voz pasiva y generar una estructura sintáctica regular de la forma sujeto-verbo-predicado. Para facilitar la recodificación de eventos derivados de voz pasiva, se sugiere incluir el prefijo `99` al inicio del código del verbo (ver sección 4.3). Este prefijo puede ayudar a los usuarios a identificar este tipo de estructuras gramaticales y recodificar el evento de forma numérica como `202051 → 99104 → 601060`, facilitando así su interpretación textual como `“tropas militares” → “arrestaron” → “miembro del crimen organizado.”`

6.3 Agregación y eliminación de duplicados

EVENTUS ID fue diseñado para identificar eventos discretos a partir del texto fuente. Con base en información del corpus, el programa recurre a un conjunto de diccionarios para identificar la fuente, acción y objetivo del evento, así como la fecha y lugar de su ocurrencia. Dada la irregularidad temporal y geográfica de la ocurrencia de eventos, el producto de codificación final genera una lista de códigos discretos que no están del todo listos para realizar análisis estadístico o visualización de datos. Para analizar los datos de manera sistemática, los usuarios deben agregar los datos de eventos codificados por EVENTUS ID en intervalos regulares de tiempo (T) y en unidades espaciales específicas (N) para conformar estructuras de datos panel (T x N).

El programa de análisis estadístico STATA ofrece procedimientos fáciles de usar para agregar datos en diferentes unidades espaciales y temporales (ver el comando `collapse` en `http://www.stata.com/help.cgi?collapse`). Dependiendo de las necesidades de cada proyecto, los usuarios pueden contar, sumar o promediar los eventos codificados.

Es importante mencionar que el uso de diversas fuentes de información suele generar múltiples reportes de un mismo evento, lo cual puede provocar la inflación artificial del número de sucesos. Como señalan algunos autores (Davenport y Ball 2002, Davenport 2009), los medios de información tienden a sobre-reportar eventos relevantes mientras que sucesos menos prominentes usualmente reciben poca atención mediática. Esto induce el riesgo de sesgo sistemático en la medición de eventos. Otra posible causa de inflación artificial de eventos radica en el hecho que las notas de prensa suelen mencionar varias veces la situación descrita en el texto. En dicho caso, la repetición y la redundancia en la narrativa periodística pueden hacer que EVENTUS ID codifique el mismo evento varias veces.

Para reducir el riesgo de inflar artificialmente el número de eventos debido al sobre-reportaje en la narrativa, los usuarios deben considerar la eliminación de eventos codificados múltiples veces. La recomendación habitual es conservar un solo evento por día en la unidad

espacial más pequeña (e.g. municipio) y eliminar los eventos repetidos en la misma unidad-día. El comando `duplicates` de STATA ofrece una forma sencilla para identificar, reportar, listar y eliminar observaciones repetidas (ver `http://www.stata.com/support/faqs/data-management/duplicate-observations/`).

7. EJEMPLO DE CODIFICACIÓN

Con base en la investigación de Osorio (2013), esta sección muestra la aplicación de EVENTUS ID para codificar eventos de violencia relacionada con el crimen organizado en México. El objetivo de dicho estudio consiste en identificar tendencias en el comportamiento de grupos criminales a partir de su interacción con actores gubernamentales y otras organizaciones delictivas. Para ello, el protocolo de codificación identifica un conjunto de eventos referentes a la aplicación de la ley por parte del estado en contra de grupos criminales, así como la reacción de las asociaciones delictivas en contra del estado y los eventos de violencia entre organizaciones criminales rivales.

El corpus de texto contiene una colección de más de 41,000 comunicados de prensa y notas de periódico provenientes de 105 fuentes de información recopiladas entre el 1 de enero del 2000 y 31 de diciembre de 2010. Las fuentes de información incluyen cuatro secretarías gubernamentales a nivel federal, 32 agencias de gobierno estatal, 11 periódicos nacionales y 58 periódicos locales.

Por una parte, el diccionario de actores considera una amplia lista de actores gubernamentales tales como instituciones, funcionarios y fuerzas del orden a nivel federal, estatal y municipal. El listado de actores también incluye un número considerable de organizaciones criminales, así como el nombre de sus líderes. Además se incluye un listado detallado de varios tipos de drogas, armamento, vehículos y propiedades. Por otra parte, el diccionario de verbos considera un amplio menú de tácticas violentas y no violentas de imposición de la ley ejercidas por agentes gubernamentales. El diccionario también incluye una lista detallada de acciones violentas perpetradas por miembros de organizaciones criminales. Finalmente, los diccionarios de lugares recopilan el nombre de todos los estados y municipios del país. Para evitar problemas de ambigüedad en la geo-referenciación de eventos, el diccionario de filtros considera un conjunto de excepciones para la identificación de lugares.

El proceso de codificación de eventos generó más de 250,000 eventos geo-referenciados a nivel municipal. Finalmente, el proceso de agregación de eventos en formato de serie de tiempo transversal cubre todos los municipios del país (N=2,456) con información diaria (T=4,017) para generar una base de datos de más de 9.8 millones de observaciones. La Figura 2 muestra las principales tendencias identificadas en la base de datos con eventos agregados a nivel nacional de manera mensual. Por una parte, el Panel A se refiere al conjunto de eventos

violentos. La serie de tiempo de *violencia estatal* muestra episodios en los que el estado ejerció acción letal en contra de presuntos miembros de organizaciones criminales. La variable *competencia* indica los eventos violentos entre grupos criminales rivales. Finalmente, la serie *represalia* indica los ataques perpetrados por grupos criminales en contra de las autoridades gubernamentales. Por otra parte, el Panel B muestra el conjunto de tácticas no violentas de aplicación de la ley incluyendo arrestos, decomisos de bienes, drogas y armamento.

[Insertar aquí Figura 2]

Finalmente, aprovechando el carácter geo-referenciado de los datos, la Figura 3 muestra los focos rojos de competencia violenta entre organizaciones criminales. El mapa fue construido mediante el análisis de Funciones de Densidad Kernel, las cuales asignan mayor elevación a las áreas que concentran mayor densidad de eventos acumulados entre 2000 y 2010 . Los datos geo-referenciados permiten identificar diferentes patrones de dispersión del conflicto. El mapa muestra zonas de alta intensidad de violencia concentrada en territorios reducidos, así como áreas donde la conflictividad es moderada pero cubre extensiones importantes del territorio.

[Insertar aquí Figura 3]

La base de datos desarrollada por Osorio (2013) muestra la capacidad de EVENTUS ID para codificar eventos a partir de información escrita en español. Los datos ofrecen información granulada acerca de una gran variedad de actores ejerciendo diferentes tipos de acciones en contra de otros actores. La información desagregada por tipo de actor y de acción a nivel diario y a escala municipal ofrece un nivel de detalle sin precedente en el análisis cuantitativo de las dinámicas de violencia relacionada con el crimen organizado en México. Estos datos son sólo una primera muestra del potencial de EVENTUS ID para la generación de información sistemática y de alto valor agregado que permita el avance de diversas agendas de investigación en América Latina.

BIBLIOGRAFÍA

Adcock, Robert y David Collier. 2001. “Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *The American Political Science Review* 95(3):529-546.

Alcoba Rueda, Santiago. 1983. “El presente de los titulares de prensa: no deíctico, protiempo anafófico." URL: http://dfe.uab.es/dfeblog/salcoba/files/2008/10/presente_titulares_tiempo_anafora.pdf

Best, Rebecca, Christine Carpino y Mark J.C. Crescenzi. 2013. “An analysis of the TABARI coding system”, *Conflict Management and Peace Science*, 30(4):335-348.

Beieler, John. 2013. “Mapping Protest Data." URL: http://johnbeieler.org/blog/2013/07/03/mapping-protest-data/ Accesado el 21 de mayo de 2014.

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. New York: Wiley.

Collier, David y Henry E Brady. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Maryland: Rowman & Littlefield Publishers.

Davenport, Christian. 2009. *Media Bias, Perspective, and State Repression: The Black Panther Party*. New York: Cambridge University Press.

Davenport, Christian y Patrick Ball. 2002. “Views to a kill: exploring the implications of source selection in the case of Guatemalan State Terror, 1977-1995." *Journal of Conflict Resolution* 46(3):427-450.

Goertz, Gary. 2005. *Social Science Concepts: A User's Guid*e. Princeton, New Jersey: Princeton University Press.

Grimmer, Justin y Brandon M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis*. 1, 1-31.

Guízar García, Elizabeth. 2004. El uso de los verbos en los titulares de cinco diarios de la ciudad de México: análisis sintáctico. Tesis doctoral. Universidad Nacional Autónoma de México, México.

Hanna, Alex. 2014. “Assessing GDELT with handcoded protest data." URL: http://badhessian.org/2014/02/assessing-gdelt-with-handcoded-protest-data/ Accesado el 21 de mayo de 2014.

King, Gary, Robert O. Keohane y Sidney Verba. 1994. *Designing Social Inquiry*. Princeton University Press.

Leetaru, Kalev y Philip A. Schrodt. 2013. “GDELT: Global Data on Events, Location and Tone, 1979-2012." URL http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf Accesado el 21 de mayo de 2014.

Lewis, M. Paul, Gary F. Simons y Fennig Charles D. 2013. “Summary by language size." URL: http://www.ethnologue.com/statistics/size Accesado el 21 de mayo de 2014.

Martínez, Francisco, Lucas Miguel y Cristian Vázquez. 2004. “La titulación en la prensa gráfica." URL: http://www.perio.unlp.edu.ar/grafica1/htmls/apuntescatedra/apunte_titulacion.pdf Accesado el 21 de mayo de 2014.

Moore, Will H. 2014a. “No More Fountains of Youth/Pots o' Gold: Conceptualization and Events Data (Part 1)." URL: http://willopines.wordpress.com/2014/03/03/no-more-fountains-of-youthpots-o-gold-conceptualization-and-events-data-part-1/ Accesado el 21 de mayo de 2014.

Moore, Will H. 2014b. “No More Fountains of Youth/Pots o' Gold: Conceptualization and Events Data (Part 2)." URL: http://willopines.wordpress.com/2014/03/04/no-more-fountains-of-youthpots-o-gold-conceptualization-and-events-data-part-2/ Accesado el 21 de mayo de 2014.

Nadal Palazón, Juan. 2009. *El Discurso Ajeno en los Titulares de la Prensas Mexicana*. México City: Universidad Nacional Autónoma de México.

O´Brien, Sean O. 2010. “Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research”. *International Studies Review*. 12: 87-104.

Osorio, Javier. 2013. “Hobbes on Drugs: Understanding Drug Violence in Mexico." Tesis doctoral. University of Notre Dame.

Osorio, Javier y Alejandro Reyes. 2014. “Eventus ID: Supervised Event Coding from Text Written in Spanish." Version 2.0. URL: http://www.javierosorio.net/#!software/cqbi Accesado el 21 de mayo de 2014.

Osorio, Javier y Alejandro Reyes. 2014a. “Web 2 Eventus." Version 2.0. URL: http://www.javierosorio.net/#!software/cqbi Accesado el 21 de mayo de 2014.

Osorio, Javier y Alejandro Reyes. 2014b. “Web Text Downloader." Version 2.0. URL: http://www.javierosorio.net/#!software/cqbi Accesado el 21 de mayo de 2014.

Price, Megan y Anita Gohdes. 2014. “Searching for Trends: Analyzing Patterns in Conflict Violence Data." URL: http://politicalviolenceataglance.org/2014/04/02/searching-for-trends-analyzing-patterns-in-conflict-violence-data/ Accesado el 21 de mayo de 2014.

Schrodt, Philip A. 2014. “TABARI. Textual Analysis by Augmented Replacement Instructions.". Version 0.8.4. URL

http://eventdata.parusanalytics.com/tabari.dir/TABARI.0.8.4b2.manual.pdf Accesado el 21 de mayo de 2014.

Schrodt, Philip A. y Deborah Gerner. 1994. “Validity Assessment of a Machine-Coded Event Data Set for the Middle East, 1982-1992." *American Journal of Political Science* 38:825-854.

Schrodt, Philip A. y Deborah Gerner. 2012. “Fundamentals of Machine Coding”. In *Analyzing International Event Data: A Handbook of Computer-Based Techniques*. URL: http://parusanalytics.com/eventdata/papers.dir/AIED.Preface.pdf Accesado el 21 de mayo de 2014.

Steinert-Threlkeld, Zachary. 2014. “Machine Coded Events Data and Hand-Coded Data." URL: http://politicalviolenceataglance.org/2014/03/19/machine-coded-events-data-and-hand-coded-data/ Accesado el 21 de mayo de 2014.

Subrahmanian, V. S. 2013, *Handbook of Computational Approaches to Counterterrorism*, Springer, New York.

Ulfedler, Jay. 2013a. “Road-Testing GDELT as a Resource for Monitoring Atrocities." URL: https://dartthrowingchimp.wordpress.com/2013/05/02/road-testing-gdelt-as-a-resource-for-monitoring-atrocities/ Accesado el 21 de mayo de 2014.

Ulfedler, Jay. 2013b. “The Future of Political Science Just Showed Up." URL: https://dartthrowingchimp.wordpress.com/2013/04/10/the-future-of-political-science-just-showed-up/ Accesado el 21 de mayo de 2014.

United Nations Global Pulse. 2012. "Big Data for Development: Opportunities & Challenges,” United Nations. URL: http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf Accesado el 21 de mayo de 2014.

Ward, Michael, Andreas Beger, Josh Cutler, Matthew Dickenson, Cassy Dorff y Ben Radford. 2013. “Comparing GDELT and ICEWS Event Data." URL: http://mdwardlab.com/sites/default/files/GDELTICEWS_0.pdf Accesado el 21 de mayo de 2014.

Weller, Nicholas y Kenneth McCubbins. 2014. “Raining on the Parade: Some Cautions Regarding the Global Database of Events, Language and Tone Dataset." URL: http://politicalviolenceataglance.org/2014/02/20/raining-on-the-parade-some-cautions-regarding-the-global-database-of-events-language-and-tone-dataset/ Accesado el 21 de mayo de 2014.

Figura 1. Proceso de codificación de EVENTUS ID.

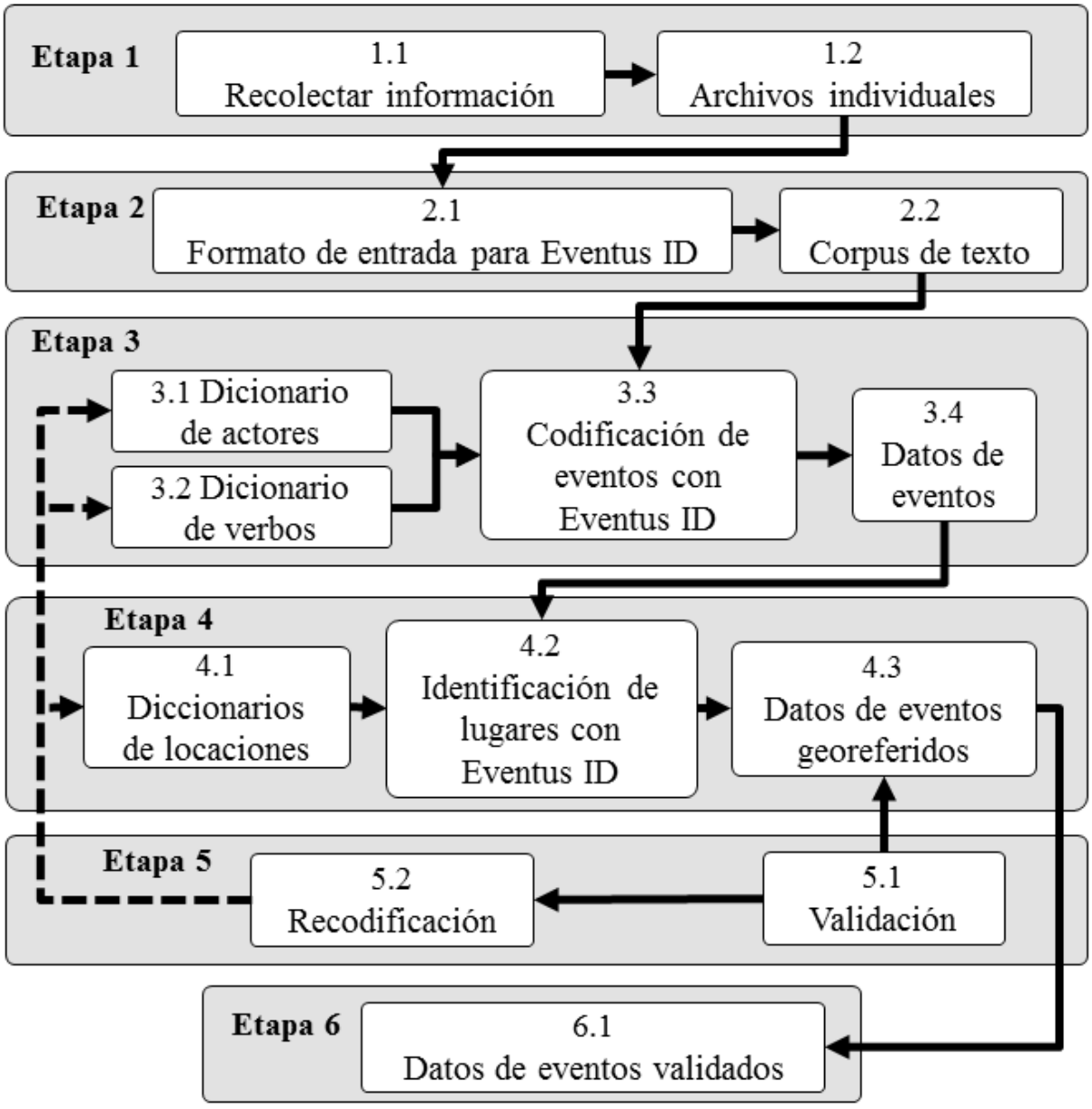


Tabla 1. Ejemplo de conjugaciones verbales en inglés y español.

Persona	Indicativo		Subjuntivo		Gerundio	Voz pasiva pasada
	Presente	Pasado	Presente	Imperfecto		
Inglés						
I	arrest	arrested	arrest	arrested	arresting	was arrested
You	arrest	arrested	arrest	arrested	arresting	were arrested
He, She	arrests	arrested	arrests	arrested	arresting	were arrested
We	arrest	arrested	arrest	arrested	arresting	were arrested
They	arrest	arrested	arrest	arrested	arresting	were arrested
Español						
Yo	arresto	Arresté	arreste	arrestara o arrestase	arrestando	fui arrestado
Tú	arrestas	arrestaste	arrestes	arrestaras o arrestases	arrestando	fuiste arrestado
Ella, él, usted	arresta	arrestó	arreste	arrestara o arrestase	arrestando	fue arrestada o fue arrestado
Nosotros	arrestamos	arrestamos	arrestemos	arrestáramos o arrestásemos	arrestando	fuimos arrestados
Vosotros	arrestáis	arrestasteis	arrestéis	arrestarais o arrestaseis	arrestando	fuisteis arrestados
Ellas, ellos, ustedes	arrestan	arrestaron	arresten	arrestaran o arrestasen	arrestando	fueron arrestadas o fueron arrestados
Vos	arrestás	arrestaste	arrestes	arrestaras o arrestases	arrestando	fuisteis arrestado

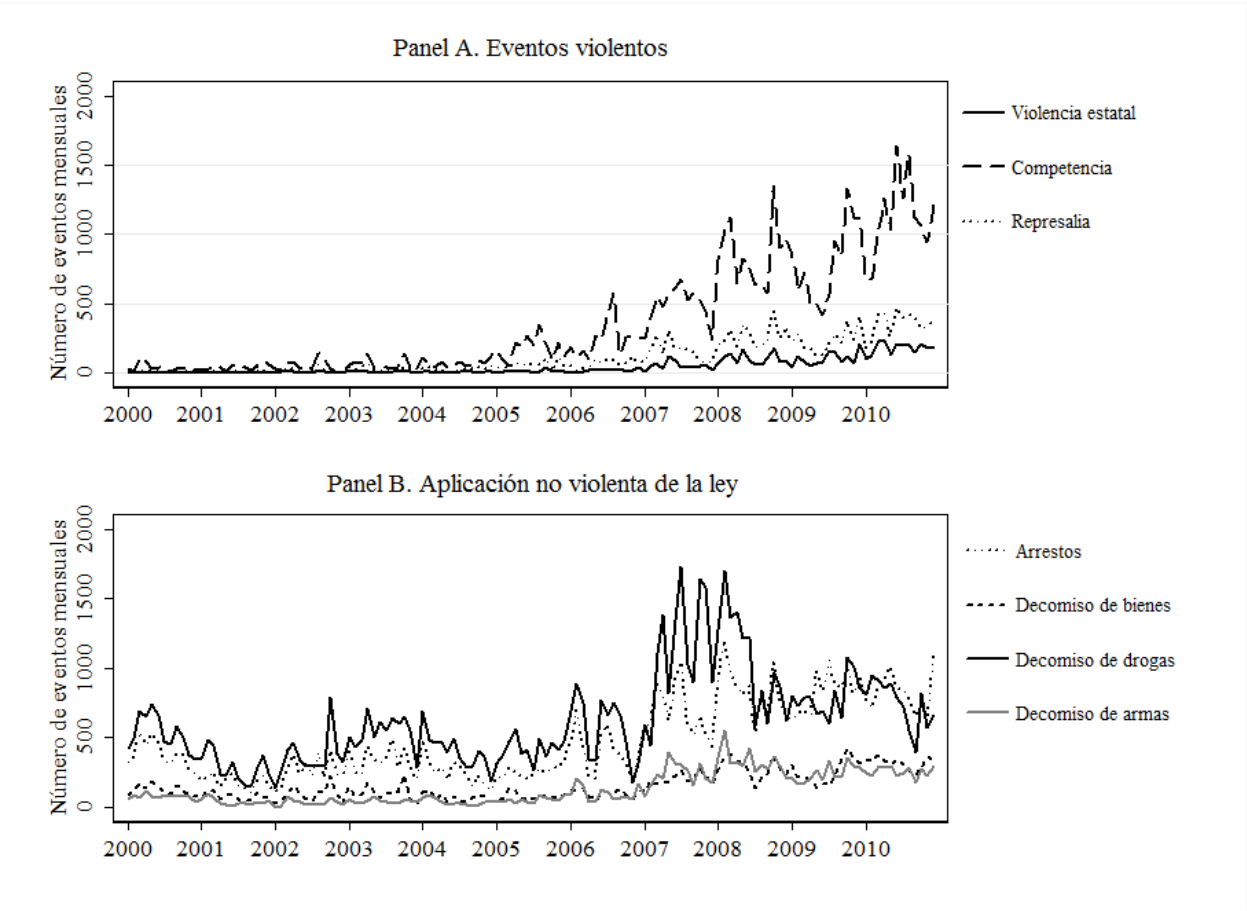
Tabla 2. Algoritmos de codificación de eventos

Algoritmo general	Algoritmo parcial
1) Búsqueda de actor fuente	1) Búsqueda de verbo
• Cargar el diccionario de actores • Iniciar la lectura del texto • Buscar primero el actor más largo • Almacenar el código actor1 cuando el actor es identificado • Pausa la búsqueda cuando el actor es detectado	• Cargar el diccionario de verbos • Iniciar la lectura del texto • Buscar primero el verbo más largo • Almacenar el código verbo cuando el verbo es identificado • Pausar la búsqueda cuando el verbo es detectado
2) Búsqueda de verbo	2) Búsqueda de actor objetivo
• Cargar el diccionario de verbos • Reiniciar la lectura a partir de la pausa anterior • Continuar leyendo el texto • Buscar primero el verbo más largo • Almacenar el código verbo cuando el verbo es identificado • Pausar la búsqueda cuando el verbo es detectado • Si no se detecta un verbo, ir al paso 4.	• Cargar el diccionario de actores • Reiniciar la lectura a partir de la pausa anterior • Continuar leyendo el texto • Buscar primero el actor más largo • Almacenar el código actor2 cuando el actor es identificado • Pausa la búsqueda cuando el actor es detectado • Si no detecta un actor, ir a paso 3.
3) Búsqueda de actor objetivo	3) Almacenar el evento
• Recargar el diccionario de actores • Reiniciar la lectura a partir de la pausa anterior • Continuar leyendo el texto • Buscar primero el actor más largo • Almacenar el código actor2 cuando el actor es identificado • Pausa la búsqueda cuando el actor es detectado	• Guardar los códigos [---] [verbo] [actor2] en la base de datos • Si no detectó un actor, entonces guardar el evento como [---] [verbo] [---] • Iniciar de nuevo en el paso 1.
4) Almacenar el evento	
• Guardar los códigos [actor1] [verbo] [actor2] en la base de datos • Si no se detectó un verbo, guardar el evento como [actor1] [---] [---] • Iniciar de nuevo en el paso 1.	

Tabla 3. Algoritmo de identificación de lugares de EVENTUS ID.

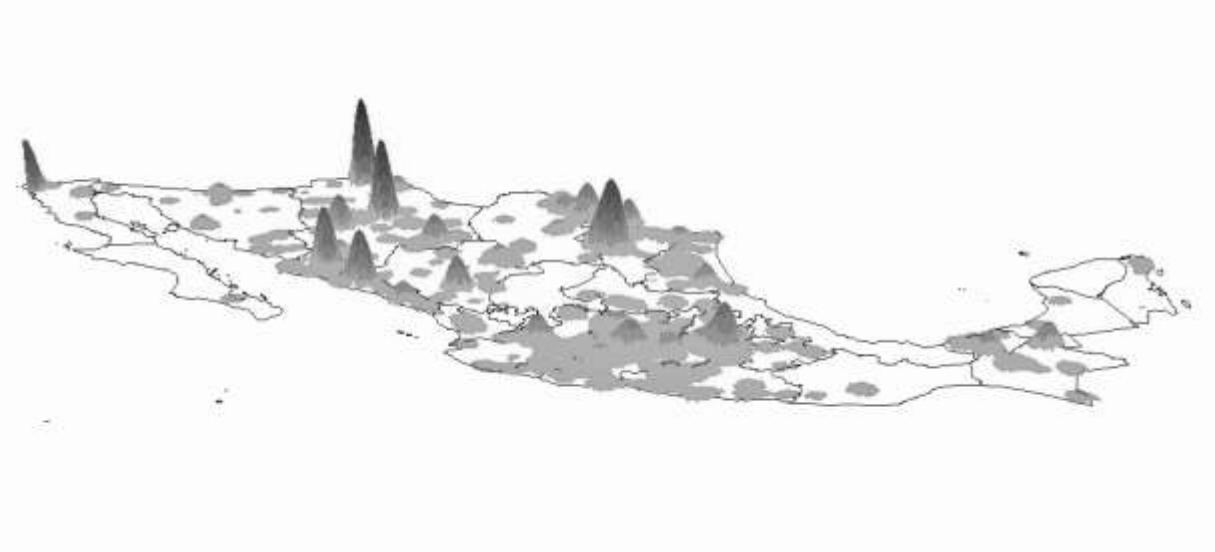
1) Identifica el evento
• Cargar la base de datos de eventos • Seleccionar un evento en una nueva línea • Identificar el párrafo específico (`NombreArchivo_P1_P2`) del que fue extraído el evento • Usar el identificador del párrafo como criterio de búsqueda
2) Identificar el párrafo en el corpus de texto
• Cargar el corpus de texto • Buscar el párrafo específico que contiene el evento identificado
3) Buscar la ubicación del evento en el párrafo fuente
• Cargar los diccionarios de localidades (estados y municipios) • Utilizar los elementos de lis diccionarios de localidades como criterios de búsqueda • Iniciar la búsqueda de lugares en el párrafo fuente • Si la ubicación es identificada, guardar el código del lugar • Sigue buscando lugares y guardando su código hasta que termine el párrafo • Si no hay más lugares en el párrafo, vaya al Paso 5 • Si no se identificaron lugares en el párrafo fuente, vaya al Paso 4
4) Expandir la búsqueda al resto del documento fuente
• Seleccionar los demás párrafos del documento fuente (`NombreArchivo`) • Busque la ubicación del evento en todos los párrafos del documento • Inicie la búsqueda por el primer párrafo • Si el lugar es identificado en el documento, impute la acción en el párrafo del evento, guarde el código y vaya al Paso 5 • Si la ubicación no es identificada en el documento, pare la búsqueda y vaya al Paso 1
5) Filtre la locación
• Cargue el diccionario de filtros • Verifique que la ubicación no corresponda a los elementos del diccionario de filtros que deben ser descartados • Si el lugar detectado empata con el contenido del diccionario de filtros, regrese al Paso 3 • Si la ubicación identificada no empata con el listado de filtros, vaya al Paso 6
6) Guarde la ubicación del evento
• Guarde los códigos de las locaciones detectadas al final de la línea del evento analizado • Inicie de nuevo en el Paso 1

Figura 2. Eventos de violencia relacionada con el crimen organizado en México.



Fuente: Osorio (2013).

Figura 3. Focos rojos de competencia violenta entre organizaciones criminales en México.



Fuente: Osorio (2013).