Guías de anotación de información de salud protegida

Plan de impulso de las Tecnologías del Lenguaje

Enrique Mota[1], Nelson Martín[1], Ángel Moreno[2], Elvira Ferrete[2], Jesús Santamaría[3], Montserrat Marimon[4], Ander Intxaurrondo[4], Aitor González-Agirre[4], Marta Villegas[4], Martin Krallinger[3,4]

1 Indizen Technologies

2 Hospital Universitario "12 de Octubre"

3 Centro Nacional de Investigaciones Oncológicas

4 Centro Nacional de Supercomputación

10 - 2018

Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para el Avance Digital, que no comparte necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores.

ÍNDICE

RESUMEN

Este documento describe las directrices para realizar el proceso de desidentificación en informes clínicos.

1 INTRODUCCIÓN

El Plan de Impulso de las Tecnologías del Lenguaje (Plan TL) tiene como objetivo fomentar el desarrollo del Procesamiento del Lenguaje Natural (PLN) y la Traducción Automática (TA) en lengua española y lenguas co-oficiales. Para ello, el Plan TL define medidas que:

- Aumenten el número, calidad y disponibilidad de las infraestructuras lingüísticas en español y lenguas co-oficiales;
- Impulsen la Industria del lenguaje fomentando la transferencia de conocimiento entre el sector investigador y la industria; e
- Incorporen a la Administración como impulsor del sector de PLN.

Uno de los objetivos del proyecto es poner a disposición de la comunidad científica y la industria un corpus biomédico exhaustivo y con licencia abierta que permita ejecutar tareas de PLN sobre grandes volúmenes de datos (big data) y replicar los experimentos.

Para múltiples tareas de PLN destinadas a la investigación científica, a procesos de análisis de datos, ciencia de datos, etc. se requiere de la desidentificación de registros médicos, (usaremos el término desidentificación como sinónimo de anonimización a lo largo de este documento).

El Centro Nacional de Investigaciones Oncológicas (CNIO) requiere que los registros médicos del paciente tengan eliminada toda la información de identificación de personas con el fin de proteger la privacidad del paciente. Según el *Health Insurance Portability and Accountability Act* (HIPAA) hay 18 categorías de identificadores de Información de Salud Protegida (ISP) del paciente, de sus familiares, del centro asistencial,... que deben ser eliminados para que un registro que se considere desidentificado.

Con esta guía se pretende establecer las bases para realizar el proceso de desidentificación del corpus del CNIO correspondiente a un total de 1.000 documentos clínicos. Este corpus será desidentificado por un grupo de 2 personas del Hospital Universitario 12 de Octubre que aplicarán las pautas marcadas por el HIPAA utilizando doble anotación seguida de rondas de comprobación de coherencia y de revisión y corrección.

Con el fin de desidentificar los registros se utiliza la herramienta Annotate[1], cada archivo tiene la ISP marcada de manera que se puede quitar/reemplazar más tarde. Esto se hará mediante la interfaz gráfica de la herramienta, y a toda la ISP se le dará una etiqueta XML que indica su categoría y tipo, en su caso.

Las anotaciones resultantes se usarán tanto para desidentificar los datos como para establecer el *gold standard* para las tareas de desidentificación de informes médicos de forma que toda la información de salud que sea privada será anotada para que pueda reemplazar con información "dummy" de forma automática.

2 REGLAS DE ANOTACIÓN MANUAL

Nuestras pautas de anotación se guiarán por los objetivos del proyecto. Haremos una interpretación lo más fiable posible a las pautas HIPAA ampliando y adaptando algunas ISPs para que se ajusten a la realidad de los registros de salud de España pero siempre desde una postura de "aversión al riesgo" dada la gran variabilidad de usuarios que tienen que interpretar la información basada en la desindentificación.

Básicamente, adaptaremos y en su caso ampliaremos la definición de la ISP 18 (cualquier número de identificación único, característica o código) para incluir otra información que está indirectamente relacionada con los pacientes y que podría usarse, por sí misma o en combinación, para identificar a los pacientes. Estos identificadores indirectos incluyen información sobre hospitales, médicos y enfermeras, así como las profesiones del paciente.

Los 18 identificadores específicos recogidos en el ISP del HIPAA son:

1. Nombres
2. Datos geográficos
3. Todos los elementos de las fechas
4. Números telefónicos
5. Números FAX
6. Correos electrónicos
7. Números de Seguridad Social
8. Números de registros médicos
9. Números de beneficiarios del plan de salud[2]
10. Números de cuenta
11. Certificado / números de licencia
12. Identificadores de vehículos y números de serie, incluidas placas
13. Identificadores de dispositivo y números de serie

[1] http://temu.bsc.es/ANONIMIZACION/annotate

[2] Este identificador no se aplica en España.

14. URL web
15. Direcciones de protocolo de Internet
16. Identificadores biométricos (es decir, escaneo retiniano, huellas dactilares)
17. Fotos de cara completa e imágenes comparables
18. Cualquier número de identificación único, característica o código

También anotamos todas las partes de las fechas, incluidos los años, así como todas las ubicaciones, incluidos las ciudades, regiones y países.

Otra ISP guiada por nuestra "aversión al riesgo" será la edad. Todas las edades en los registros deben ajustarse en consecuencia, de modo que la edad de la persona no se pueda calcular a partir de la información en otros documentos. Anotar todas las edades, nos permite modificar fácilmente todas las edades en los registros de una persona. Este enfoque implica cambios menores a la precisión médica, pero proteger las identidades del paciente que es lo más importante para este proyecto, esto nos permite crear un corpus de ISP que podría servir como datos de entrenamiento para los sistemas que traten de abordar la desidentificación automática.

La aversión al riesgo es una consideración clave en nuestras pautas de anotación.

También tenemos en cuenta que, en última instancia, este corpus se utilizará para la investigación de técnicas de PLN y, por lo tanto, es necesario representar con precisión los registros clínicos reales. Por lo tanto, la ISP en este corpus necesita preservar su semántica a través de la desidentificación y los procesos de generación sustitutos. Este objetivo nos lleva a definir categorías de ISP que son muy precisas. En el siguiente apartado se muestran los criterios de anotación.

Las etiquetas y los criterios de anotación que usaremos en este proyecto serán los que se exponen a continuación.

Las reglas de anotación se han clasificado en 5 tipos:

- Reglas generales (Reglas-G): reglas positivas y negativas que se aplican a todas las etiquetas de menciones, incluyen reglas ortográficas generales.
- Reglas de clases entre etiquetas (Regla-C): reglas que definen a qué clase pertenece una mención.
- Reglas positivas (Reglas-P): reglas que especifican los ISP que se deben anotar.
- Reglas negativas (Reglas-N): reglas que especifican los ISP que no hay que anotar.
- Reglas multipalabra (Regla-M): reglas que especifican si un grupo de palabras debe anotarse bajo una única etiqueta o no.

Adicionalmente se incluyen diversos ejemplos para cada regla en los que la evidencia textual se resalta en gris. Las reglas negativas. además, mostrarán en rojo el texto que no se debe anotar.

2.1 REGLAS GENERALES (REGLAS-G)

- **G1. Solo anotar la información que necesita ser reemplazada cuando se re-identifique el archivo.**

- **G2. No anotar los términos o palabras clave mismas que se refieren a las etiquetas de las ISP, por ejemplo, "nombre", "edad", "sexo", etc. ya que por sí solos estas palabras o términos no constituyen información sensible.**

  Ejemplos:

"*Nombre*: **Rafael…"**	**[correcto]**	**no se etiqueta** *"Nombre"*
"Nombre: Rafael…"	**[incorrecto]**	**se etiqueta** *"Nombre"*

- **G3. No incluir en la anotación manual los espacios ni signos de puntuación que aparezcan antes o después de cada mención.**

  Ejemplos:

"Nombre: **Rafael…"**	**[correcto]**	**no se incluyen ni espacios en blanco ni signos de puntuación**
"Nombre: **Rafael…"**	**[incorrecto]**	**se incluye espacio en blanco y signo de puntuación antes de la mención**
" Miguel Fresnos, …"	**[incorrecto]**	**se incluye espacio en blanco y signo de puntuación después de la mención**

- **G4. Anotar las menciones, aunque tengan errores tipográficos u de otro tipo.**

2.2 REGLAS DE CLASES ENTRE ETIQUETAS (REGLAS-C)

- **C1. ID_SUJETO ASISTENCIA vs. OTROS_SUJETO_ASISTENCIA**

Anotar los grupos poblacionales de MESH, como grupo continental o étnico (Anexo II), como identificadores de paciente.

Ejemplos: Paciente de origen subsahariano

Paciente musulmán varón

- **C2. NUMERO_TELEFONO vs. FAX**

Anotar los números de teléfono que vayan precedidos por el campo FAX con la etiqueta FAX.

Ejemplos: … calle Manuel Luna 32. CP 02080 Albacete. FAX: 637 249 823 ...

2.3 REGLAS ESPECÍFICAS DE ETIQUETAS DE CLASES DE MENCIONES (REGLAS-P, REGLAS-N, REGLAS-M)

En este apartado se incluyen reglas positivas (Reglas-P), reglas negativas (Reglas-N) y reglas multipalabra (Reglas-M) para cada uno de los ISP.

2.3.1 NOMBRE_SUJETO_ASISTENCIA

Reglas positivas

- **P1.1. Anotar el nombre propio y todos los apellidos del paciente.**

  Ejemplos: Nombre: Rafael
  Apellidos: Calvo Martín

- **P1.2. Anotar las abreviaturas y las iniciales, incluso aquellas que no parecen coincidir con un nombre.**

  Ejemplos: Nombre: Dña. M. Jesús
  Nombre: Fco. José

- **P1.3. Anotar los apodos, motes, alias, sobrenombres, hipocorísticos.**

  Ejemplos: [INV][3]Nombre: M. del Mar (Marita)
  [INV]Apellidos: Vinardell
  [INV]Apellidos: esposa del Sr. Alvarado
  [INV]Ernesto "Che" Guevara

- **P1.4. Anotar los títulos nobiliarios.**

  Ejemplos: [INV]Nombre: Duque de Alba...

- **P1.5. Anotar el plural de los antropónimos**

  Ejemplos: [INV]Nombre: acudieron los Pérez para ...

[3] Todos los ejemplos marcados con [INV], son inventados. No representan en ningún caso datos extraídos de un documento clínico real.

Reglas negativas

- **N1.1. NO incluir en la etiqueta los tratamientos, por ejemplo, Sr., Sra., Dña., etc.**

  Ejemplos: Nombre: Dña. M. Jesús

Reglas multipalabra

- **M1.1. Anotar como una sola mención los nombres y apellidos de pacientes si aparecen seguidos en el texto separados sólo por espacios o por guiones.**

  Ejemplos: Nombre: Francisco Javier

  Apellidos: Martínez-Aguado

- **M1.2. Anotar como una sola mención los nombres y apellidos de pacientes, aunque alguno de los dos parezca dudoso, si no se tiene información para desambiguar.**

  Ejemplos: [INV]… francisco musulmán de 50 años de edad ...

2.3.2 EDAD_SUJETO_ASISTENCIA

Reglas positivas

- **P2.1. Anotar todas las edades del paciente, no solo las de más de 90. Estas pueden aparecer como valor de la etiqueta "edad" o en alguno de los apartados de la historia clínica del paciente.**

  Ejemplos: Edad: 68

  Antecedentes: Paciente de 76 años

  Historia actual: Enfermo varón de 40 años

- **P2.2. Anotar también las edades cuando estas sean inferiores al año, es decir, cuando las unidades de medida temporal sean *horas*, *días, semanas o meses*. Incluir en la etiqueta las unidades de medida temporal.**

  Ejemplos: Recién nacido de tres días de vida

  Edad: 3 días de nacido

- **P2.3. Incluir en la etiqueta las unidades de medida temporal: *año(s), mese(s), semana(s), día(s), hora(s).***

  Ejemplos: Recién nacido de tres días de vida

  Historia actual: Enfermo varón de 40 años

2.3.3 SEXO_SUJETO_ASISTENCIA

Reglas positivas

- **P3.1. Anotar todas las variantes que hacen referencia al sexo del paciente, por ejemplo, *hombre, varón, mujer, niño/a*. Estas pueden aparecer como valor de la etiqueta "*sexo*" o en alguno de los apartados de la historia clínica del paciente.**

  Ejemplos: Enfermo varón

  Antecedentes: Mujer sometida a una…

  Antecedentes: Niña de 11 años…

  Historia actual: Paciente masculino negro de 39 años…

  Sexo: Hombre

- **P3.2. Anotar también las abreviaturas "M", "H".**

  Ejemplos: Sexo: M

Reglas negativas

- **N3.1. NO anotar los tratamientos, por ejemplo, Sr., Sra., Dña., etc.**

  Ejemplos: Nombre: Dña. M. Jesús

Reglas multipalabra

- **M3.1. Anotar como una sola mención si aparecen varios seguidos, por ejemplo, por error.**

  Ejemplos: [INV]Sexo: M H

2.3.4 FAMILIARES_SUJETO_ASISTENCIA

- **P4.1. Anotar el nombre propio, todos los apellidos y los apodos, motes, alias o sobrenombres de los familiares del paciente si son mencionados**

  Ejemplos: [INV]Acude al centro asistencial acompañado de su esposa M. Jesús Calvo

- **P4.2. Anotar el número, el sexo, el grado de parentesco, etc. si son mencionados**

  Ejemplos: [INV]… casado con dos hijos y una hija …

  Ejemplos: [INV]… acuden con él dos primos y su tía …

- **P4.3. Anotar las edades de los familiares del paciente si son mencionadas.**

  Ejemplos: Personales: N. de Bolivia, en España desde el 2003. Según refiere, Tiene un hija de 14 años de apo última mujer y Otros hijo que vive en USA.

2.3.5 NOMBRE_PERSONAL_SANITARIO

Reglas positivas

- **P5.1. Anotar el nombre propio y los apellidos de doctor o personal sanitario.**

  Ejemplos: Médico: Josep Rubio Palau

- **P5.2. Anotar también las abreviaturas y las iniciales.**

  Ejemplos: Médico: J.R Palau

Reglas negativas

- **N5.1. NO anotar la abreviatura de tratamiento "Dr./ Dra."**

  Ejemplos: Remitido por: Dr. Pablo Garrido Abad

- **N5.2. NO anotar el número de registro de profesionales.**

  Ejemplos: Médico: Josep Rubio Palau NºCol: 08-08-25574

Reglas multipalabra

- **M5.1. Anotar como una sola mención los nombres y apellidos de personal sanitario si aparecen seguidos en el texto separados sólo por espacios o por guiones.**

  Ejemplos: Remitido por: Dr. Josep Rubio Palau

- **M5.2. Anotar como una sola mención los nombres y apellidos de pacientes, aunque alguno de los dos parezca dudoso, si no se tiene información para desambiguar.**

  Ejemplos: [INV]... el médico francisco musulmán de 50 años de edad …

2.3.6 FECHAS

Reglas positivas

- **P6.1. Anotar todas las fechas, incluidas las fechas en forma de texto sin formato.**

  Ejemplos: Fecha de ingreso: 20/09/2016

  Fecha de nacimiento: 23-octubre-19724

Fecha de nacimiento: 22 Septiembre de 2018

- **P6.2. Anotar cualquier fecha del calendario, incluidos años, estaciones, meses y días festivos.**

  Ejemplos: Otoño del 2017

  En el mes de Marzo de ese mismo año…

  ...se le diagnostica a principios del año 2004 un carcinoma…

  … diagnosticada en 2014…

  se normalizó en el plazo de 4 meses (Junio 04).

- **P6.3. Anotar los días de la semana.**

  Ejemplos: ... el jueves 2 de Agosto se le diagnosticó ...

- **P6.4. Si la frase tiene 's (es decir,' 'en los años 90), anotar '' 90s ".**

  Ejemplos: ... en los 90s tuvo un accidente …

- **P6.5. Anotar las palabras o frases que hagan referencia a una fecha conocida (festividades, santos, etc.).**

  Ejemplos: [INV]... en nochevieja ingresó …

  [INV]se le dio el alta en navidad ...

  [INV]acudió a urgencias el día de la madre ...

  [INV]en san juan acudió a su centro de salud ...

  [INV]ingresó disfrazado (era halloween) en ...

Reglas negativas

- **N6.1. NO incluir la hora del día.**

  Ejemplos: ... el jueves 2 de Agosto a las 14:41 ingresó en el hospital ...

Reglas multipalabra

- **M6.1. NO anotar dentro de una única etiqueta los intervalos de tiempo "*desde/de… hasta/a…*"; cada una de las fechas se anotará por separado.**

  Ejemplos: había precisado hemodiálisis desde 1980 a 1983

2.3.7 PROFESIÓN

Reglas positivas

- **P7.1. Anotar cualquier trabajo que aparezca en el texto.**

Ejemplos: trabajador de la construcción

Reglas negativas

- **N7.1. NO anotar modificadores, por ejemplo, *jubilado* o *trabajador.***

  Ejemplos: trabajador de la construcción jubilado

Reglas multipalabra

- **M7.1. Anotar como una sola mención las profesiones que estén compuestas por más de una palabra.**

  Ejemplos: … es un ingeniero informático con …

2.3.8 HOSPITAL

Reglas positivas

- **P8.1. Anotar todos los nombres de los hospitales o centros asistenciales.**

  Ejemplos: Hospital Universitario de La Princesa
  Hospital San Juan de la Cruz

- **P8.2. Incluir en la etiqueta los modificadores que aparecen junto con el nombre del hospital, como *complejo* y *hospital* (ver ejemplos en R7.1).**

  Ejemplos: Servicio del Aparato digestivo Complejo Hospitalario de Navarra

Reglas multipalabra

- **M8.1. En caso de duda, anotar de forma separada los nombres de hospitales de cualquier grupo de palabras que le preceda o le siga.**

  Ejemplos: HU 12 de octubre

2.3.9 ID_CENTRO DE SALUD

Reglas positivas

- **P9.1. Anotar todos los nombres de los centros de salud.**

  Ejemplos: Centro de Salud Cea Bermúdez
  Centro de Salud Infanta Mercedes

2.3.10 INSTITUCIÓN

Reglas positivas

- **P10.1. Anotar como institución el nombre de instituciones que no se refieran explícitamente a un hospital ni a un centro de salud.**

  Ejemplos: Información facilitada por el Centro Nacional de Investigaciones Oncológicas

  Juan Ortiz vicepresidente de Sanitas

- **P10.2. Anotar siglas y alias, incluso aquellas que aparezcan en minúsculas.**

  Ejemplos: Informe reportado por el CNIO

  Informe reportado por el cnio

Reglas multipalabra

- **M10.1. Anotar como una sola mención las instituciones que estén compuestas por más de una palabra.**

  Ejemplos: Información facilitada por el Centro Nacional de Investigaciones Oncológicas

2.3.11 CALLE[4]

Reglas positivas

- **P11.1. Anotar tanto el nombre de la vía como el número, piso, escalera... que aparezcan detrás o delante del nombre de la vía.**

  Ejemplos: Domicilio: Avenida Melchor Fernández 59. 4,2

  Paraje Torrecárdenas s/n 04009 Almería (España)

  Remitido por: Dr. Tomás Lázaro Rodríguez Collar Hospital Universitario Dr. Carlos J. Finlay. 114 y 31. Marianao. 11500. La Habana. (Cuba)

- **P11.2. Anotar los tipos de vías que aparecen junto con el nombre de la vía, como *avenida, calle, plaza*...**

  Ejemplos: Domicilio: Avenida Melchor Fernández 59. 4,2

  Plaza Juan Carlos I, Nº 7 03370 Redován. Alicante.

  Domicilio: Paseo Echegaray y Caballero 100, 3,1

- **P11.3. Anotar los tipos de vías que aparecen como iniciales o abreviaturas junto con el nombre de la vía, por ejemplo, *C/, avda., plza, etc.***

[4] Esta clase también incluye la denominación con nombre propio de fincas y edificios y los casos de direcciones en los que pueden aparecer nombres como: almacenes, apartamentos, distritos, pabellones, parcelas, playas, polígonos industriales, puertos, aeropuertos y estaciones y las localizaciones de carreteras.

Ejemplos: Raquel Santesteban Muruzábal C/ Irunlarrea, 3 1008 Pamplona

Jorge Subirá Ríos. AV. San Francisco 7, 3D 50006 Zaragoza. (España).

Domicilio: Av. Litoral, 30, 1C

Domicilio: Pº rosales, 28,3A

Reglas multipalabra

- **M11.1. Anotar como menciones separadas cuando haya intercalados otros ISP.**

  Ejemplos: AV. San Francisco 7, 50006 Zaragoza, tercero derecha

2.3.12 TERRITORIO

Reglas positivas

- **P12.1. Anotar los códigos postales, incluida la letra en aquellos en los que aparezca, así como las localidades (ciudades, pueblos...).**

  Ejemplos: CP: 01022

  CP: E-28015

  Localidad/ Provincia: Barcelona/Barcelona

  Madrid/Comunidad de Madrid

- **P12.2. Anotar las localidades que vayan precedidas de un domicilio.**

  Ejemplos: Calle Ginzo de limia 56, 1D Albacete

  Calle Strachan, 4- 2º piso 29015- Badajoz

- **P12.3. Anotar la comunidad autónoma, la provincia y las áreas geográficas.**

  Ejemplos: Río Júcar, s/n E-28935 Móstoles **(**Madrid)

  Localidad/ Provincia: Madrid/Comunidad de Madrid

- **P12.4. Anotar alternativas estilísticas, accidentes geográficos, regiones naturales y ecorregiones, comarcas, espacios naturales protegidos, divisiones territoriales de carácter administrativo, regiones militares, barrios y urbanizaciones**

  Ejemplos: [INV]la ciudad eterna (= Roma)

  [INV]la falla de San Andrés

  [INV]la Sierra

  [INV]la Mancha

[INV]parque nacional de Doñana

[INV]la diócesis de Cuernavaca

[INV]barrio de Lavapiés

[INV]pertenece a la urbanización torrejoncillo de los higos ...

- **P12.5. Anotar los continentes**

  Ejemplos: [INV]lEuropa

Reglas multipalabra

- **M12.1. Anotar como dos menciones separadas la ciudad y el código cuando aparezcan seguidos en el texto separados solo por espacios o por guiones.**

  Ejemplos: Calle Strachan, 4- 2º piso 29015- Badajoz

2.3.13 PAÍS

Reglas positivas (Reglas-P)

- **P13.1 Anotar todos países que aparezcan en el texto, tanto si aparecen como valor del campo *País* como si aparecen en alguno de los apartados de la historia clínica del paciente.**

  Ejemplos: País: España

  28020 Madrid, Madrid, España

2.3.14 NÚMERO_TELÉFONO

Reglas positivas

- **P14.1. Anotar todos los números de teléfono que aparezcan en el texto.**

  Ejemplos: Número de teléfono de la madre: 630 304 365

  El número de móvil de su esposa es el 633 349 565

- **P14.2. Anotar los números del buscapersonas como números de teléfono.**

  Ejemplos: 6527- 3368

2.3.15 NÚMERO_FAX

Reglas positivas

- **P15.1. Anotar los números de teléfono que aparecen como valor del campo *Fax (o FAX).***

Ejemplos: Fax: 0998 455879

2.3.16 CORREO_ELECTRÓNICO

Reglas positivas

- **P16.1 Anotar todas las direcciones de correo electrónico, aparezcan o no como valor del campo *Email* (o Correo electrónico, o cualquiera de sus variantes).**

  Ejemplos: Email: jrubiopalau@yahoo.es

2.3.17 ID_SUJETO_ASISTENCIA

Reglas positivas

- **P17.1 Anotar todos los identificadores de paciente como códigos CIPA, NHC, DNI, NIF, CIE y pasaporte.**

  Ejemplos: NHC: 987654

  CIPA: nhc-150679

  [INV]DNI: 38987678

  [INV]NIF: 38987678-P

- **P17.2. Anotar los identificadores de paciente por grupos poblacionales de MESH, como por ejemplo grupo continental o étnico (Referencias MESH grupos poblacionales en Anexo II).**

  Ejemplos: Paciente masculino negro de 39 años…

  Paciente musulmán varón

- **P17.3. Anotar las orientaciones sexuales: "homosexual", "heterosexual", "transexual", etc.**

  Ejemplos: Varón homosexual de 21 años sin antecedentes patológicos ...

Reglas negativas

- **N17.1. NO anotar el prefijo 'nhc-' del código CIPA.**

  Ejemplos: CIPA: nhc-150679

- **N17.2. NO anotar el sexo ya que tiene su propia categoría.**

  Ejemplos: Paciente masculino negro de 39 años…

- **N17.3. NO anotar como grupo étnico si hay dudas sobre si lo es o forma parte de otro ISP (como, por ejemplo, su apellido).**

Ejemplos: ... francisco musulmán de 50 años de edad ...

2.3.18 ID_CONTACTO_ASISTENCIAL

Reglas positivas

- **P18.1. Anotar el número de identificación del episodio, contacto, problema, proceso...**

  Ejemplos: Episodio:1520368541

2.3.19 ID_ASEGURAMIENTO

Reglas positivas

- **P19.1. Anotar el número de la afiliación a la seguridad social (NASS).**

  Ejemplos: NASS: 33 4568642 23

2.3.20 ID_TITULACIÓN_PERSONAL_SANITARIO

Reglas positivas

- **P20.1. Anotar el número de colegiado del personal sanitario.**

  Ejemplos: NºCol: 08-08-25574

2.3.21 ID_EMPLEO_PERSONAL_SANITARIO

Reglas positivas

- **P21.1 Anotar el número de empleado en el centro hospitalario.**

  Ejemplos: Nº empleado: u000999; e999000

2.3.22 IDENTIF_VEHÍCULOS_NRSERIE_PLACAS

Reglas positivas

- **P22.1 Anotar datos de identificación de elementos que, por pertenecer o estar relacionados con ellos, permiten la identificación de los individuos como los identificadores de vehículos, matrículas, números de bastidor de vehículos (sí se mantienen los nombres de marcas y modelos).**

  Cabe señalar que en el caso de las matrículas de los coches, se consideran información sensible todas las matrículas, aunque pertenezcan a personas jurídicas.

  Ejemplos: Matricula: 5478GDV

Nº bastidor: VF1RFD00653635032

2.3.23 IDENTIF_DISPOSITIVOS_NRSERIE

Reglas positivas

- **P23.1 Anotar las direcciones IP, MAC ADDRESS.**

  Ejemplos: ip: 192.168.0.23

  mac: 00:0a:95:9d:68:16

2.3.24 DIREC_PROT_INTERNET

Reglas positivas

- **P24.1 Anotar las direcciones de protocolo de internet incluyendo TCP, HTTP, SMTP, FTP…**

  Ejemplos: https://www.sample.com

  ftp://wwwsample.com

2.3.25 URL_WEB

Reglas positivas

- **P25.1. Anotar las URLs, direcciones WEB.**

  Ejemplos: url: www.sample.com

2.3.26 IDENTIF_BIOMÉTRICOS

Reglas positivas

- **P26.1 Anotar otros identificadores biométricos, como por ejemplo, escaneo retiniano, huellas dactilares…**

  Ejemplos: [INV]... ha sido necesario acceder a su identificador de huella dactilar FP537WXZ391 para poder identificar al paciente ...

2.3.27 NUMERO_IDENTIF

Reglas positivas

- **P27.1 Anotar cualquier otro tipo de identificación no incluido en las otras categorías, como el número de identificación de socio, del carnet de la biblioteca...**

Ejemplos: NºSocio: 5547889955

2.3.28 OTROS_SUJETO_ASISTENCIA

Reglas positivas

- **P28.1 Anotar la información que no podía clasificarse como cualquier otra información médica protegida, pero que aún podría proporcionar información sobre el paciente. Esta clase incluye aquellos datos de identificación de elementos que, por pertenecer o estar relacionados con ellos, permiten la identificación de los individuos como números de cuentas corrientes, datos registrales (generalmente registro de la propiedad o mercantil) o notariales (generalmente números de protocolo), números de liquidaciones o números de expedientes.**
  Esta clase también incluye a personas jurídicas que permitan la identificación del sujeto, tales como el nombre de la empresa para la que trabaja.

  Ejemplos: [INV]tiene un tatuaje del pato Donald en la mano derecha …
  [INV]madre Teresa Rodríguez
  [INV]Síndrome de Down
  [INV]el jefe de Gobierno …
  [INV]la ministra de Defensa

Las categorías detalladas abordan dos objetivos adicionales:

En primer lugar, podríamos ser expansivos para marcar cualquier ISP potencial, y se ajustará el alcance de la tarea según sea necesario en el futuro. Por ejemplo, para fines de depuración de datos, queríamos adherirnos a nuestra estrategia de aversión al riesgo y marcar toda la ISP potencial; sin embargo, para la evaluación de la tarea compartida de desidentificación, podríamos seleccionar un subconjunto de la ISP y enfocarnos en las categorías que le importa a HIPAA.

En segundo lugar, otros usuarios del corpus podrían enfocarse en categorías de ISP específicas para su investigación a partir de sus objetivos de desidentificación. Por ejemplo, para la ISP NOMBRE, mantenemos las distinciones entre los NOMBRE_SUJETO_ASIDTENCIA y NOMBRE_PERSONAL_SANITARIO porque algunos proyectos de desidentificación no incluyen los nombres de los trabajadores del hospital en la ISP. Aquí, NOMBRE_PERSONAL_SANITARIO se utiliza como un término general para todo el personal del hospital, incluidas enfermeras, farmacéuticos, recepcionistas, etc.

Un tercer objetivo de la categorización de la ISP de grano fino es permitirnos recopilar la ISP para un tratamiento uniforme mientras mantenemos sus categorías de grano fino simplificando el proceso de generación de información sustituta. Por ejemplo, podemos reunir subconjuntos de categorías HIPAA 4-17 en CONTACTO e ID. La mayoría de estas categorías de HIPAA son simplemente cadenas alfanuméricas y se tratan de manera similar para la generación sustituta. La disponibilidad de categorías de ISP de grano fino permite que este paso se lleve a cabo de manera eficiente, simplificando el proceso de generación sustituta sin ninguna pérdida de semántica.

3 PROCEDIMIENTO DE ANOTACIÓN

3.1 DESIDENTIFICACIÓN

Aplicamos las pautas expuestas anteriormente en las tareas compartidas de desidentificación de 1000 documentos provistos por el CNIO. Los registros longitudinales de cada paciente en nuestro corpus son anotados con la ISP por dos anotadores independientes trabajando en paralelo. A los dos anotadores se les asigna aleatoriamente un conjunto de registros de pacientes diferente. Después de eso, utilizamos múltiples controles para asegurar que no se filtre ISP, como se describe a continuación. En [1] también se utilizó un método de "doble anotación"; sin embargo, en [2] se utilizaron tres anotadores independientes, y en [3] se utilizó la anotación en serie. Como se mencionó anteriormente, durante el proceso de anotación, los autores realizaron un estudio para determinar si la anotación en paralelo o en serie funcionaba mejor para capturar toda la ISP en un registro; los resultados de este estudio mostraron que ninguno de los métodos fue más efectivo que el otro [4].

Para el software de anotación, usamos el Entorno de anotación multipropósito suministrado por el CNIO.

Cada archivo que se desidentifique será revisado por dos anotadores. Toda información que cumpla con los criterios de ISP debe etiquetarse con la etiqueta de anotación adecuada, y luego debe indicarse el tipo de ISP cuando corresponda.

3.2 GENERACIÓN SUSTITUTA Y VERIFICACIÓN FINAL DE ISP

Antes de que podamos poner los registros médicos a disposición de los interesados, necesitamos ocultar toda la ISP reemplazándola con sustitutos realistas. Se crea un software automático de generación sustituta para realizar el reemplazo. Una descripción completa del proceso de generación sustituta y sus complejidades se encuentran en [5]; resumimos el proceso en el resto de esta sección. Tratar todos los registros de un paciente como un solo archivo nos facilita mantener la continuidad entre los registros durante la generación sustituta, de modo que todos los nombres

en el registro médico longitudinal del paciente son reemplazados consistentemente con el mismo sustituto, las fechas son todas compensadas por la misma cantidad, etc.

Para cada documento, cambiamos todas las ISP de FECHAS por el mismo número aleatorio de años, meses y días. Para los ISP NOMBRE, para cada documento nuevo asignamos aleatoriamente cada letra del alfabeto a otra letra, y decidimos previamente que todos los NOMBRE comenzando con, por ejemplo, la letra A serían reemplazados por un sustituto que comenzará con G como una forma de simplificar la generación inicial para NOMBRE. Luego, prestamos atención para mantener la información de género y para reemplazar a los NOMBRE con NOMBRE del género apropiado al seleccionarlos de las listas generadas a partir de los datos del censo. Por ejemplo, asumiendo un mapeo de A a G, y de F a D, '' Angie Ferrerro '' se podría convertir en '' Grace Dollard ''. Para preservar la información de correferencia, todas las ocurrencias de un NOMBRE auténtico serán reemplazadas por el mismo sustituto. Los sustitutos imitan la estructura de la ocurrencia original mapeando ''A. Ferrerro", ''Sra. Ferrerro", "Angie" a "G. Dollard", "Sra. Dollard" y "Grace", respectivamente. Seguimos el mismo procedimiento para generar sustitutos para CALLE, TERRITORIO y PROFESIÓN, aunque sin las asignaciones alfabéticas. Usamos una lista precompilada de sustitutos, de la cual seleccionamos según corresponda, mientras conservamos la información de la correferencia.

Modificamos todos los números, incluidos NÚMERO_TELÉFONO, NÚMERO_FAX y todas las subcategorías de ID, seleccionando al azar nuevas cadenas de dígitos / letras de la misma longitud y formato.

Cualquier otra ISP, como CORREO_ELECTRÓNICO, URL_WEB, se reemplazan inicialmente por cadenas de caracteres aleatorios; tras ejecutar el software generador de sustitutos, se podrán modificar según sea necesario para hacerlas más realistas.

En el proceso se modifican otros dos tipos de sustitutos manualmente: fechas ambiguas como 02/03, que podría ser el 3 de febrero, el 2 de marzo o febrero de 2003; y sobrenombres, o faltas de ortografía de NOMBRE y cualquier otra ISP.

El encargado de generar la sustitución revisará de forma manual que todas las ISP han sido sustituidas después de la generación sustituta automática. También normaliza la ISP sustituta, por ejemplo, verificando que la terminación apropiada (a/o) aparece frente a un sustituto y modificando los sustitutos según sea necesario (por ejemplo, ''camboyano'' podría ser reemplazado por el suplente ''china'', que se cambia a ''chino'' para que coincida con el texto circundante). También se recrearán errores ortográficos y otros errores en los datos indirectos: por ejemplo, si un paciente llamado ''Marissa'' es identificado como ''Marrisa'' en un registro diferente, los sustitutos de los datos identificados podrían haber sido cambiados a ''Sara'' y ''Sarrah''.

4 ANEXO I - INFORMACIÓN DE APOYO

DEPARTAMENTOS HOSPITALARIOS GENÉRICOS

Unidad de agudos

Cardiología

Unidad de cuidados coronarios / UCC

Cuidado crítico

Otorrino (OTO)

Urgencias / URG

Sala Urgencias

Gastroenterología

Cirugía General

Unidad de cuidados intensivos geriatricos

Ginecología

Hematología

Unidad de cuidados intensivos (UCI)

Medicina Interna

Maternidad

Departamento de registros médicos

Unidad neonatal

Unidad de cuidados intensivos neonatales (UCIN)

Nefrología

Neurología

Obstetricia

Terapia ocupacional

Oncología

Sala de operaciones

Oftalmología

Ortopedia

Unidad de cuidados intensivos pediátricos (UCIP)

Farmacia

Terapia física

Unidad de cuidados postanestésicos

Unidad de Psiquiatría / Psiquiatría

Radiología

Reumatología

Cirugía

Atención de urgencias

Urología

5 ANEXO II - DOCUMENTACIÓN DE REFERENCIA

Documento base para el proyecto:

https://www.sciencedirect.com/science/article/pii/S1532046415001823

Aplicación:

Anotador1: http://temu.bsc.es/ANONIMIZACION/annotate/

Anotador2: http://temu.bsc.es/ANONIMIZACION/annotate-2/

Resultados:

Anotador1(izda), anotador2(dcha): http://temu.bsc.es/ANONIMIZACION/brat/diff.xhtml?diff=%2Fannotator-1-ann%2F#/annotator-2-ann/es-S0004-06142005000900013-1

Anotaciones solapadas: http://temu.bsc.es/ANONIMIZACION/brat/index.xhtml#/mix/es-S0004-06142005000900013-1

Plan de trabajo:

https://docs.google.com/spreadsheets/d/1tu4cfXHCND0NQjpaYq6j-Cdb5XvGIQsmvpTjAuLBukw/edit?userstoinvite=jesussantamaria505@gmail.com&ts=5b28a2c5&actionButton=1#gid=1537094442

Otras guías de anotación:

- Guidelines for Annotating Personal Identifiers in the Clinical Text Repository of the National Institutes of Health https://scrubber.nlm.nih.gov/annotation/pdf.papers/Guidelines.2016.06.28.pdf

- Guías de anotación y documentación general sobre NER en inglés.
    - guías de anotación de ACE: https://drive.google.com/open?id=1gViscaNwyS1VFpXWMfhhn6dkvpZ8u_Fk
    - presentación NER:

https://drive.google.com/open?id=1CByQMgEwvnTq6ATFCtu5b4G_R8r7BBL_

- presentación de-identificación:

https://drive.google.com/open?id=1yKSusQDijNdkmSABxg8rFiKkDbmahmKS

- algunos enlaces a guías on-line:

https://drive.google.com/open?id=10lUXJzOChEo3bcFbW_Xp8t3yX-uBqORroDiFoVfZc_0

- **Referencias Mesh:**

Características Pacientes: https://www.ncbi.nlm.nih.gov/mesh/1000077

Grupos de población, grupos étnicos: https://www.ncbi.nlm.nih.gov/mesh/68044382

Buscar conceptos equivalentes español: https://babelmesh.nlm.nih.gov/index_spa.php?com=

6 REFERENCIAS

[1] L. Deleger, T. Lingren, Y. Ni, M. Kaiser, L. Stoutenborough, K. Marsolo, M. Kouril, K. Molnar, I. SoltiPreparing an annotated gold standard corpus to share with extramural investigators for de-identification research J. Biomed. Inform., 50 (2014), pp. 173-183, **10.1016/j.jbi.2014.01.014**

[2] I. Neamatullah, M. Douglass, L.H. Lehman, A. Reisner, M. Villarroel, W.J. Long, P. Szolovits, G.B. Moody, R.G. Mark, G.D. CliffordAutomated de-identification of free-text medical recordsBMC Med. Inform. Decis. Mak., 8 (2008), p. 32**, 10.1186/1472-6947-8-32**

[3] Ö. Uzuner, Y. Luo, P. SzolovitsEvaluating the state-of-the-art in automatic de-identificationJ. Med. Inform. Assoc., 14 (5) (2007), pp. 550-563, **10.1197/jamia.M2444**

[4] A. Stubbs, Ö. Uzuner, De-identification of medical records through annotation, in: Nancy Ide, James Pustejovsky (Eds.), Chapter in Handbook of Linguistic Annotation, Springer, 2015

[5] A. Stubbs, Ö. Uzuner, C. Kotfila, I. Goldstein, P. SzolovitzChallenges in synthesizing replacements for PHI in narrative EMRs Aris Gkoulalas-Divanis, Grigorios Loukides (Eds.), Chapter in Medical Data Privacy Handbook, Springer, Anticipated Publication (2015)

7 GLOSARIO DE SIGLAS Y ACRÓNIMOS

CNIO	Centro Nacional de Investigaciones Oncológicas
HIPAA	Health Insurance Portability and Accountability Act

ISP	Información de Salud Protegida
Plan TL	Plan de Impulso de las Tecnologías del Lenguaje
PLN	Procesamiento de Lenguaje Natural
TA	Traducción Automática