“Arquitectura de Referencia para Anonimizar Documentos”

Anexo A: Relevamiento de Arquitecturas de Anonimización

Ing. Horacio Vico
Tutor: MSc. Ing. Daniel Calegari
Montevideo - 2013.

Índice

Índice de figuras

Índice de tablas

1. Introducción

El objetivo de la tesis es estudiar el problema de la anonimización en el contexto de los sistemas de gestión documental. Particularmente se desean abordar en profundidad las técnicas y tecnologías disponibles para despersonalizar y anonimizar documentos en forma automática, y proponer una arquitectura de referencia para un software de anonimización.

Este documento presenta un relevamiento de las técnicas existentes en anonimización de documentos, que incluye el estudio de diversas propuestas arquitecturales y herramientas de software vinculadas.
El documento se organiza de la siguiente forma:

En la Sección 2 - Gestión Documental, se introducen conceptos generales vinculados con el tema de éste trabajo de investigación, aportando información contextual que se considera relevante. Se brindarán algunas definiciones de conceptos que serán utilizados frecuentemente a lo largo del trabajo.

En la Sección 3 - Marcos de Aplicación, se estudian distintos ámbitos de aplicación de la anonimización.

En la Sección 4 - Arquitecturas de Anonimización, se presentan distintas propuestas de arquitecturas que surgen de distintos proyectos o artículos académicos estudiados.

En la Sección 5 - Instanciación tecnológica de los módulos, se investigan herramientas de software que permiten instanciar los distintos componentes y módulos estudiados en la Sección 4.

2. Gestión Documental

En esta sección se introducirán varios conceptos que serán de uso frecuente a lo largo del trabajo, por estar fuertemente vinculados con la temática central del mismo (la anonimización). En primer lugar introduciremos el concepto de Gestión Documental.

Algunas definiciones de Gestión Documental:

- "La gestión documental consiste en el uso de la tecnología y procedimientos que permiten la gestión y el acceso unificado a información generada en una organización." [45]
- "Conjunto de normas técnicas y prácticas usadas para administrar el flujo de documentos de todo tipo en una organización, permitir la recuperación de información desde ellos, determinar el tiempo que los documentos deben guardarse, eliminar los que ya no sirven y asegurar la conservación indefinida de los documentos más valiosos, aplicando principios de racionalización y economía." [7]
- "Área de gestión responsable de un control eficaz y sistemático de la creación, la recepción, el mantenimiento, el uso y la disposición de documentos de archivo, incluidos los procesos para incorporar y mantener en forma de documentos la información y prueba de las actividades y operaciones de la organización."[44]
- "Conjunto de actividades administrativas y técnicas tendientes a la planificación, manejo y organización de la documentación producida y recibida por las entidades, desde su origen hasta su destino final, con el objeto de facilitar su utilización y conservación" [45]

El tema de estudio de este trabajo, la anonimización, surge como una actividad más de la gestión documental, y de allí que éste concepto resulta relevante de definir.

Un componente frecuente asociado a la gestión documentales son las bases de datos específicas que se utilizan en éste contexto, es decir las bases de datos documentales. A continuación se presenta una mejor definición de éstas.

Bases de datos documentales, SGBDD y Enterprise Content Management Las bases documentales son bases de datos especializadas en el almacenamiento de documentos. Una de sus características fundamentales es que permiten la indexación a texto completo, y a través de este proceso facilitan la realización de consultas y búsquedas más potentes que las que se pueden realizar sobre una base de datos tradicional[31]. En las bases de datos documentales el registro está asociado a un documento, pudiendo ser este un documento electrónico, una imagen, un gráfico o cualquier contenido audiovisual.

Por sobre los motores de Bases de Datos Documentales, se construyen habitualmente sistemas que denominaremos Sistemas de Gestión de Bases de Datos Documentales (SGBDD). Se trata de herramientas de software orientadas a la

gestión del ciclo de vida de los documentos almacenados en las bases de datos documentales, con foco especial en requerimientos de búsqueda de la información contenida en ellos.

Se puede definir un SGBDD como "...un sistema computarizado, un conjunto de programas, utilizado para rastrear y almacenar documentos electrónicos y/o imágenes de documentos soportados en papel", que ofrece las siguientes funcionalidades [45] :

- Capacidad para almacenar información textual en forma estructurada
- Capacidad para manejar información textual de longitud grande y variable Capacidad para recuperar con rapidez, en base a la generación de índices, registros que responden a un criterio de búsqueda
- Capacidad para realizar búsquedas multicriterio utilizando la lógica booleana.
- Capacidad para administrar tesauros y diccionarios terminológicos.

Actualmente se incorpora un concepto adicional en la gestión documental, denominado "Enterprise Content Management" (EMS), el cual se considera la evolución de los SGBDD. Según la "Association for Information and Image Management", Enterprise Content Management (ECM) son las estrategias, métodos y herramientas utilizadas para capturar, manejar, salvaguardar, preservar y entregar contenido y documentos relacionados con procesos organizacionales. ECM cubre la gestión de la información dentro del ámbito de una organización, ya sea que esa información esté en forma de documento de papel, un archivo electrónico, una base de datos impresa e incluso un email. Como se expresó anteriormente, es considerada la "evolución" de los sistemas de gestión documental, y lo que provee es una formalización de procedimientos adecuados para organizar y almacenar documentos en una organización, y otro contenido que se pueda relacionar con los procesos de negocio de la organización. El término EMS engloba estrategias, métodos y herramientas que se utilizan durante el ciclo de vida del contenido.

A continuación se presentará el concepto central que motiva éste trabajo, la anonimización, tema que surge directamente de la implementación bases de datos documentales y los SGBDD descritos anteriormente dentro de organizaciones, cuando las mismas manejan información personal o confidencial.

3. Anonimización y despersonalización

Con la incorporación de la gestión documental, los procesos y tecnologías descritos en la sección anterior como herramientas de uso frecuente en el seno de las organizaciones, son numerosas las bases de datos documentales en poder de éstas. Dependiendo de su naturaleza y el dominio del negocio de la organización, estas bases de datos muchas veces son fuente de conocimiento que pueden ser aprovechadas por otras organizaciones o personas, en ocasiones con un enorme beneficio científico, técnico o social.

Sin embargo muchas veces los documentos almacenados en estas fuentes de información, contienen información personal o sensible de personas físicas o jurídicas, cuya privacidad debe ser garantizada por la organización que gestiona la base documental. Existen fundamentos normativos y legislativos que hacen que la protección de los datos personales se vuelva una responsabilidad plausible de sanciones civiles o penales.

La anonimización y la despersonalización de documentos, resulta una tarea tediosa o posiblemente inmanejable cuando se trata de bases documentales muy grandes, y es allí donde aportan un enorme valor técnicas que permitan identificar en forma automática o asistida éstos datos sensibles. En nuestro país existe legislación reciente vinculada a la protección de datos personales. También se están realizando esfuerzos importantes en incorporar tecnologías de la información en los procesos de gobierno, con una creciente disponibilidad de documentos públicos hacia los ciudadanos. Esto hace que la anonimización automática de documentos pueda ser una tecnología aplicable en el medio local en diferentes ámbitos, y resulta por tanto un tópico interesante para desarrollar en profundidad.

El verbo anonimizar es de reciente aceptación en el idioma español. Tal es así que el último diccionario de la Real Academia publicado (22a edición, año 2001) no lo define. Sin embargo en su diccionario en línea ya se lo ha incorporado aclarando que será incluido en la siguiente edición del diccionario (23a edición). La R.A.E. define anonimizar como "expresar un dato relativo a entidades o personas, eliminando la referencia a su identidad." [14]

Una muy buena definición de anonimización, se da en una ley española [8] que regula la investigación biomédica, dominio en el cual se estudiará ésta temática más adelante. La ley 14/2007 del Reino de España, define en su artículo tercero la anonimización como el "proceso por el cual deja de ser posible establecer por medios razonables el nexo entre un dato y el sujeto al que se refiere".

Existen además dos niveles de "anonimización". La despersonalización (de-identification) y la anonimización propiamente dicha (anonimization). La anonimización implica la quita irreversible de toda información que permita identificar a un individuo u organización. La despersonalización sin embargo añade la posibilidad de que se guarde algún registro referencial que permita a una entidad autorizada o de confianza acceder a los datos personales eliminados. La ley española citada, también da dos definiciones en relación a los datos que son anonimizados o despersonalizados respectivamente.

Se define como dato anonimizado o irreversiblemente disociado, a aquel "dato

que no puede asociarse a una persona identificada o identificable por haberse destruido el nexo con toda información que identifique al sujeto, o porque dicha asociación exige un esfuerzo no razonable, entendiendo por tal el empleo de una cantidad de tiempo, gastos y trabajo desproporcionados."

Como dato codificado o reversiblemente disociado, se entiende a aquel "dato no asociado a una persona identificada o identificable por haberse sustituido o desligado la información que identifica a esa persona, utilizando un código que permita la operación inversa." Este dato fue procesado en algún procedimiento de despersonalización.

En un artículo en la web, vinculado a tecnologías semánticas y de procesamiento de lenguaje [27], se describe el proceso en forma clara.

De acuerdo a ésta definición, la anonimización implica:

- Eliminar o sustituir algunos nombres de personas (físicas o jurídicas), direcciones y demás información de contacto, números identificativos, apodos o cargos.
- Eliminar o sustituir algunos lugares mencionados (ciudades, barrios, regiones, instalaciones, monumentos, áreas naturales, etc.)
- Mantener otros nombres de entidades (personas, organizaciones o lugares) cuando aportan información relevante para el caso y no facilitan la identificación.
- En ocasiones es necesario también filtrar fechas o cantidades monetarias.
- Si se sustituyen referencias a entidades por etiquetas, es necesario mantener la consistencia a lo largo de un mismo documento, a pesar de que existan variaciones en la denominación (por ejemplo, si no se usa el nombre completo en todo el texto, si se usan alias, o si existen variantes debido a errores ortográficos).

Como se verá a continuación, existen diversos ámbitos de aplicación de la anonimización como herramienta en la gestión documental.

3.1. Marcos de Aplicación de la Anonimización

La anonimización puede resultar de utilidad en cualquier ámbito donde se generen, gestionen o se trabaje de alguna manera con bases de datos documentales, donde los documentos contienen información confidencial. Particularmente donde dichos documentos tienen un valor intrínseco de índole científica o técnica, que es independiente de la información confidencial contenida en éstos.

3.1.1. Gobierno electrónico

Los sistemas de gestión documental y las bases de datos documentales son de gran utilidad y frecuente uso en casi cualquier ámbito de la gestión gubernamental. Por naturaleza las organizaciones de índole burocrática como los gobiernos,

manejan enormes volúmenes de documentos impresos y electrónicos. Por tanto la gestión documental es una necesidad de orden para garantizar un funcionamiento fluido de los procesos de administración. En este ámbito la anonimización de documentos puede aparecer como una necesidad frecuente en dicha gestión documental, ya que por motivos de transparencia muchos documentos deberán estar accesibles al ciudadano, pero a su vez se deben proteger datos personales contenidos en ellos.

Normativa uruguaya En Uruguay, el 11 de agosto de 2008 se promulgó la Ley de Protección de Datos personales[12], que regula la responsabilidad en la protección de los datos personales en poder de personas físicas o jurídicas en el territorio nacional. El objetivo de la ley es "...establecer un marco jurídico claro y necesario para garantizar y hacer efectivo uno de los derechos fundamentales del ser humano, como es el derecho a la protección de los datos de carácter personal y por tanto de la intimidad de las personas."[19] Se crea además en este acto legislativo, la Unidad Reguladora de Datos Personales, en la órbita de AGESIC (Agencia para el Desarrollo del Gobierno Electrónico y la Sociedad de la Información y el Conocimiento).

La ley da una serie de definiciones desde el punto de vista jurídico que resultan interesantes en el contexto de éste trabajo:

Dato personal: información de cualquier tipo referida a personas físicas o jurídicas determinadas o determinables.

Se consideran datos públicos, los cuales no requieren consentimiento informado a:

- Para personas físicas: nombres y apellidos, documento de identidad, nacionalidad, domicilio y fecha de nacimiento.
- Para personas jurídicas: razón social, nombre de fantasía, registro único de contribuyentes, domicilio, teléfono e identidad de las personas a cargo de la misma.

Dato sensible: datos personales que revelen origen racial y étnico, preferencias políticas, convicciones religiosas o morales, afiliación sindical e informaciones referentes a la salud o a la vida sexual.
Base de datos: conjunto organizado de datos personales que sean objeto de tratamiento o procesamiento electrónico o no, cualquiera que fuere la modalidad de su formación almacenamiento, organización o acceso.
Tratamiento de datos: operaciones y procedimientos sistemáticos, de carácter automatizado o no, que permitan el procesamiento de datos personales, así como también su cesión a terceros a través de comunicaciones, consultas, interconexiones o transferencias.
Titular de los datos: persona cuyos datos sean objeto de un tratamiento incluido dentro del ámbito de acción de la ley.
Responsable de la base de datos o del tratamiento: persona física o jurídica, pública o privada, propietaria de la base de datos o que decida sobre la

finalidad, contenido y uso del tratamiento.
Encargado del tratamiento: persona física o jurídica, pública o privada, que sola o en conjunto con otros trate datos personales por cuenta del responsable de la base de datos o del tratamiento.
Usuario de datos: toda persona, pública o privada, trate datos, ya sea en una base de datos propia o a través de conexión con los mismos.

La ley 18.331 impone un serie de responsabilidades al poseedor de bases de datos que contengan datos personales. Se establece que los datos personales deberán ser veraces, adecuados, ecuánimes y no excesivos en función de la finalidad para la que fueron recabados. Los datos deben ser actualizados de requerirse y si caducan deben ser suprimidos (principio de veracidad). Los datos no podrán utilizarse para una finalidad distinta para la que fueron obtenidos y deberán ser eliminados si ya no son útiles para sus fines (existen excepciones por valor histórico, estadístico o científico, que permite conservar los datos). No podrán comunicarse datos entre bases de datos sin previo consentimiento informado del titular (Principio de finalidad). Según el Principio de previo consentimiento informado, el titular debe haber prestado si consentimiento libre, previo, expreso e informado el que deberá documentarse para tratar los datos.
Quedan exentos de esto los datos públicos y datos: De fuentes públicas de información, recabados para el ejercicio de funciones del Estado aquellos que provengan de una relación contractual, científica o profesional del titular de los datos y sean necesarios para su desarrollo o cumplimiento, o datos para uso exclusivo personal o domestico por personas físicas o jurídicas.

El responsable o usuario de la base de datos debe garantizar la seguridad y confidencialidad de los datos. El artículo 10 prohíbe registrar datos en bases de datos que no reúnan las condiciones técnicas de integridad y seguridad. (Principio de seguridad de los datos) Quienes hayan obtenido datos de fuentes legítimas están obligados, aun después de terminada la relación con el responsable de la base de datos, a guardar el secreto profesional, usarlos de forma reservada y exclusivamente para las operaciones habituales de su giro o actividad. Esta Ley prohíbe toda difusión de los datos a terceros. (Principio de reserva) [19]

3.1.2. Ámbito Judicial

En un esquema de separación de poderes ("de Montesquieu"), posiblemente el poder del Estado que maneja la mayor cantidad de información sensible de los ciudadanos es el Poder Judicial, que es el encargado de administrar la justicia en un Estado de Derecho. La administración de justicia por su alcance universal a la ciudadanía, muchas veces implica indagar en aspectos privados de las personas o las organizaciones, y ésta información puede quedar registrada en distintos documentos (expedientes, sumarios, presumarios, sentencias, etc). Muchas veces los litigios judiciales involucran información de terceros directa o indirectamente involucrados, cuya privacidad se debe proteger también con especial cuidado. Un ejemplo concreto es toda la información relativa a menores y adolescentes que puede manejarse en la órbita de la justicia de familia, cuando se abren causas de

divorcio, pensiones alimenticias, investigación de paternidad, e incluso violencia doméstica, donde los derechos de los menores deben ser protegidos por encima de todo.

En el ámbito judicial se pueden utilizar SGBDD para gestionar gran parte de los procesos y flujos de trabajo en la órbita jurisdiccional y administrativa. Los sistemas de gestión de expedientes cuentan con módulos de gestión documental para manejar la gran variedad de documentos y resoluciones que pueden conformar un expediente judicial.

Desde el punto de vista jurisdiccional, una aplicación concreta de SGBDD donde aparece el problema de la anonimización son los sistemas de búsqueda de jurisprudencia. La jurisprudencia se compone de las sentencias judiciales ya dictadas, las cuales pueden servir como base jurídica para fundamentar nuevas resoluciones.

En el derecho anglosajón, la jurisprudencia es vinculante, es decir que la jurisprudencia sienta un precedente que en general exige que las nuevas resoluciones mantengan el mismo sentido o espíritu que los precedentes similares. En nuestro país esto no es así, la jurisprudencia es una fuente de consulta más para los magistrados, quienes pueden perfectamente tomar una postura contraria a lo que dicta la jurisprudencia en casos similares. Sin embargo esto último no suele ser lo más común, y los jueces suelen tomar en consideración la jurisprudencia existente al tomar una decisión judicial.

En Uruguay desde el año 1998 el Poder Judicial cuenta con una base de jurisprudencia denominada Base de Datos Jaime Zudáñez (BDZ), inicialmente pensada para gestionar las sentencias de la Suprema Corte de Justicia y de consulta exclusiva de los magistrados judiciales. El alcance de este SGBDD fue extendido hace unos cinco años incorporando además la jurisprudencia de los Tribunales de Apelaciones, y pasando a denominarse "Base de Jurisprudencia Nacional" (BJN).

Desde este año la base de datos es accesible al público general a través de la web.[24]

En esta aplicación en particular, el cual el autor de éste trabajo ha participado, tanto en su especificación, definición de la arquitectura e implementación. Por tal motivo se realizará una descripción del flujo de trabajo asociado en la gestión de los documentos (sentencias), para ilustrar un caso de uso de un SGBDD con anonimización. A continuación se enumeran los pasos que sigue una sentencia hasta su publicación para acceso público.

1. Un grupo de magistrados (Ministros de Tribunales de Apelación o la Suprema Corte de Justicia), dictan una sentencia,. Dicha sentencia es firmada por los magistrados y notificadas a las partes interesadas, y a partir de ese momento inicia su ciclo en el SGBDD "Base de Jurisprudencia Nacional".

2. La sentencia es ingresada al sistema por un funcionario administrativo del Tribunal. Los magistrados le entregan un documento generado en alguna herramienta ofimática (MS Office, OpenOffice). El funcionario ingresa el texto de la sentencia al sistema, y agrega una serie de metadatos básicos,

como pueden ser el número de la sentencia, el Tribunal que la dictó, el tipo de sentencia (Definitiva o Interlocutoria), qué magistrados la firmaron y cual la redactó, las materias con las que se corresponde (Derecho Civil, Familia, Penal, Laboral).

3. Una vez se ingresa toda esta información primaria, la sentencia pasa al departamento de jurisprudencia. Este departamento es el encargado de darle metadatos más técnicos al documento. En particular se "tematiza" la sentencia, seleccionando figuras jurídicas que corresponden de una estructura de árbol predefinida. Por ejemplo, una figura sería "Derecho Civil -> Demanda por daños y perjuicios -> Accidente de tránsito". Notar que no necesariamente esta figura tiene por que ser mencionada en forma explícita en el texto de la sentencia, pero un profesional del derecho al leerla puede determinar que esta sentencia se corresponde a tal o cual figura. Por tanto se da un valor agregado importante al documento que luego es fundamental a la hora de realizar consultas a la base de datos.

4. En el departamento de jurisprudencia **se anonimizan los documentos**, es decir que se sustituyen los datos sensibles que puedan contener (nombres propios, direcciones, teléfonos), por datos genéricos.

5. Una vez se finaliza con el tratamiento de la sentencia en la órbita del departamento de jurisprudencia, la misma es publicada.

6. Estando publicada, la sentencia queda disponible para que cualquier usuario la encuentre cuando realiza consultas "textuales" y parametrizadas sobre la base. Las búsquedas se aplican por sobre el texto plano de la sentencia, así como sobre toda la información adicional y metadatos que se fueron agregando durante el proceso de tratamiento de la misma.

El proceso de anonimización de sentencias en el Poder Judicial uruguayo se realiza en forma manual, y resulta un proceso bastante tedioso. El usuario debe leer la sentencia, buscando nombres propios y otros datos sensibles, e irlos sustituyendo por identificadores genéricos. Por ejemplo, se sustituye el nombre de una persona por AA y la de otra persona por BB, y luego se utiliza ese sustitución a lo largo de toda la sentencia. La única herramienta que se provee es un "buscar y sustituir" propio de cualquier editor de textos. Las sentencias se ingresan como se mostró desde archivos Word u OpenOffice, y la información sensible no viene señalada de manera alguna. Al almacenarse en el sistema se guarda la sentencia en formato HTML. En definitiva, para incorporar un procedimiento de anonimización automático o semiautomático se debería trabajar con texto no estructurado.

3.1.3. Ciencias biomédicas

En el ámbito médico se generan gran cantidad de documentos con información relevante para consulta profesional. El ejemplo más notable es el de las historias clínicas. Estos documentos guardan información médica de un paciente,

que puede ser utilizada para realizar un mejor diagnóstico o tratamiento de otros pacientes con síntomas similares. Hoy es habitual que los centros asistenciales cuenten con algún sistema de gestión médica, que entre otras cosas posee alguna base de datos documental con las historias clínicas. Sin embargo existen restricciones importantes a la hora de hacer pública esa información para consulta de otros profesionales de la medicina, ya que las historias contienen información personal y sensible de los pacientes, las cuales se encuentran habitualmente protegidas por leyes de protección de datos personales. Por tal motivo, en sistemas de gestión documental orientados al ámbito médico, un elemento a tener muy en cuenta es la anonimización de documentos.

En nuestro país en los últimos años se comenzó a legislar en relación a ésta información de los pacientes, estableciéndose normativas que aplican a la gestión de las historias clínicas y los derechos de los usuarios vinculados a dicha documentación. Tal es así que la Ley 18.335 de 2008 [11], establece que "la historia clínica es de propiedad del paciente, será reservada y sólo podrán acceder a la misma los responsables de la atención médica y el personal administrativo vinculado con éstos, el paciente o en su caso la familia el el Ministerio de Salud Pública cuando lo considere pertinente." Es decir que se establece la confidencialidad de la información de la historia clínica. Sin embargo la normativa no estudia en profundidad cuál es la información sensible o personal de las historias clínicas, sino que restringe su divulgación directamente a un grupo de personas y organizaciones específicas. El posterior decreto del Poder Ejecutivo Nº 274/010 [32] amplía la restricción estableciendo además que "los servicios de salud y los trabajadores de la salud deberán guardar reserva sobre el contenido de la historia clínica, y no podrán revelarlo a menos que fuere necesario para el tratamiento del paciente o mediare orden judicial o conforme a lo dispuesto por el Artículo 19 de la Ley Nº 18.335." Otra referencia de interés es un documento disponible en la página web de AGESIC, en la cual se mencionan las leyes y decretos que constituyen el marco legal disponible a nivel nacional en materia de historias clínicas.[17]

En una palabra, no existe legislación que permita disponer de ésta información de utilidad científica por parte de terceros, siempre protegiendo la confidencialidad del paciente. Por éste motivo se presentará el caso de Estados Unidos, donde sí existe desde hace más de quince años una ley que regula detalladamente éstos aspectos.

Ley H.I.P.A.A. en EE.UU. Los Estados Unidos cuentan con una ley denominada H.I.P.A.A. (Health Insurance Portability and Accountability Act) de la administración Clinton aprobada en 1996 [36], que regula entre otras cosas aspectos vinculados a la protección de la privacidad y la seguridad de información clínica de los individuos. Esta ley es exhaustiva detallando aquellos elementos que deben ser protegidos (anonimizados o despersonalizados).

De acuerdo al HIPAA en Estados Unidos, un centro asistencial puede utilizar los datos de un paciente para investigación sin su consentimiento, si se realiza un proceso de despersonalización de la información. Para tener una idea de

lo exhaustivo que puede ser el proceso de anonimización o despersonalización en documentos no estructurados, en el caso de HIPAA la información médica protegida (PHI por sus siglas en inglés) abarca 18 grupos de datos, entre los cuales destacan por ejemplo: nombres propios, números telefónicos, direcciones, identificadores personales, etc, etc.

H.I.P.A.A. es una norma relevante y muy vinculada al tema anonimización, ya que su aprobación y vigencia tuvo como consecuencia que la academia y la industria del software norteamericana realizaran grandes esfuerzos para aportar soluciones que permitieran procesar los grandes volúmenes de información clínica que se generan día a día. De hecho un gran cantidad de la información científica en inglés que se encuentra vinculada a anonimización, tiene alguna vinculación con H.I.P.A.A., proyectos vinculados a ésta o sistemas de gestión documental médicos.

4. Arquitecturas de Anonimización

En esta sección se estudiarán algunos modelos de arquitectura para sistemas de anonimización, que surgen de trabajos académicos y herramientas de software disponibles. El objetivo es estudiar cada una en forma individual, para luego analizar aspectos comunes, y características individuales que se logren identificar.

En primer lugar se entiende pertinente enumerar las técnicas existentes para automatizar los procesos de anonimización, ya que las mismas tienen una incidencia directa en los componentes que se verán en las distintas arquitecturas.

4.1. Arquitectura ANONIMYTEXT

Uno de artículos académicos que se encontraron que refieren a trabajos sobre documentos en idioma castellano, es el denominado ANONIMYTEXT: Anonimization of Unstructured Documents [41]. En este documento se propone una arquitectura completa para un software de anonimización de información médica no estructurada, y se utiliza una combinación de técnicas para lograr la identificación de la información sensible en el texto. Se describe a continuación someramente esta arquitectura. El sistema "Anonimytext" se compone de los siguientes subsistemas:

1. Módulo de inducción por diccionarios
2. Etiquetador ("Tagger")
3. Asesor ("Adviser")
4. Módulo para revisión por expertos
5. Anonimizador

En una primera instancia un experto del negocio (un médico) utiliza el módulo de inducción por diccionarios, para definir una lista de conceptos "sensibles" que aparecen en los documentos, por ejemplo: "nombres, direcciones, fechas". Para cada uno de ellos el experto deberá proveer una lista de ejemplos de los mismos que se utilizan como "semillas" para generar un diccionario "inducido". Un diccionario inducido es un concepto que permite, a partir de una semilla y utilizando un diccionario obtener un diccionario de "equivalencias" adaptado al dominio particular. Por ejemplo, el experto puede definir el concepto "Nombre", la semilla "Jorge" y con un diccionario que contenga (entre otras cosas) nombres propios en castellano se generará un diccionario inducido que contiene todos los nombres propios. En una siguiente etapa, el módulo etiquetador realiza un análisis semántico del texto, y etiqueta los distintos conceptos que aparecen en el mismo. Para ello puede utilizar los diccionarios inducidos generados previamente (siguiendo con el ejemplo, donde aparezca "Jorge" puede etiquetar este término como un nombre. Pero también se puede alimentar de tesauros especializados del negocio, en este caso de estudio se sugiere el UMLS

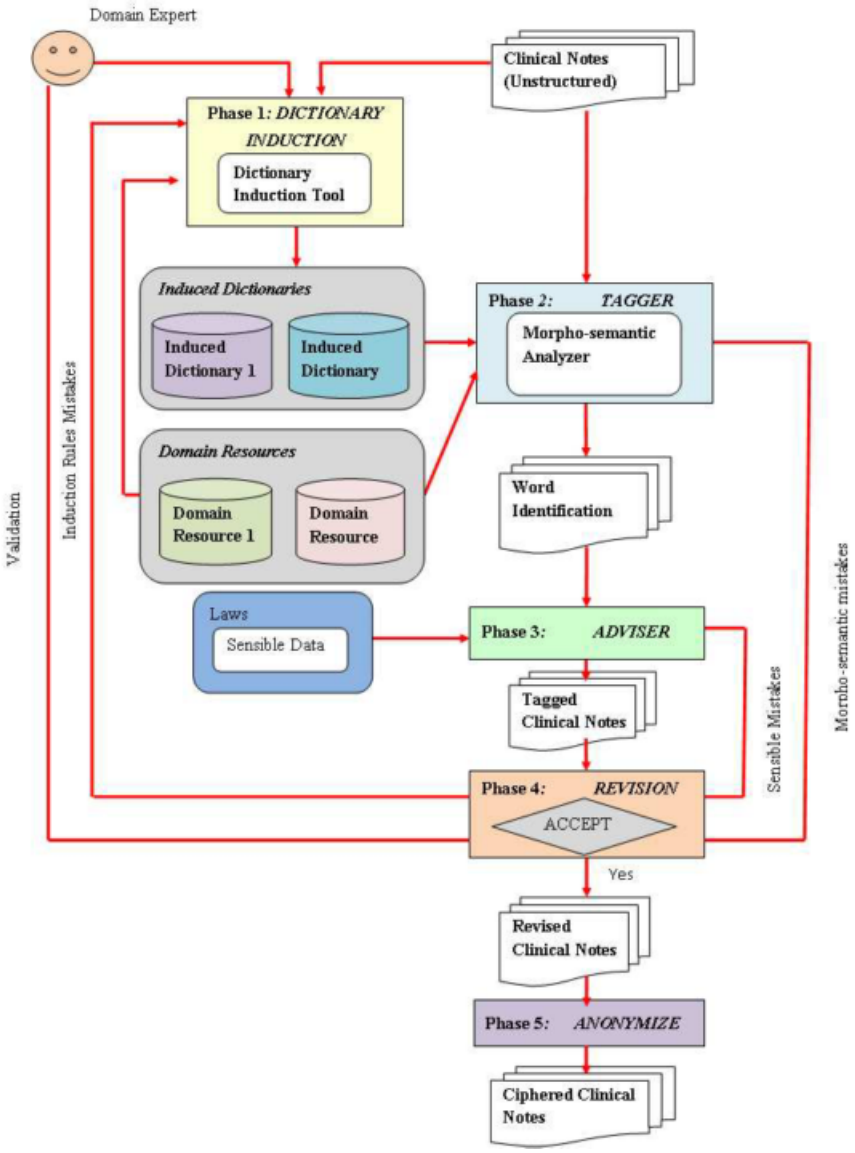

Figura 1: Arquitectura ANONIMYTEXT

Methathesarus, un compendio de millones de conceptos médicos y biomédicos. Una vez etiquetado, el texto pasa el módulo "asesor", el cual detectará la información sensible del documento basándose en la configuración que se lo provea de acuerdo a la normativa legal a aplicar (es decir, dependiendo de las reglamentaciones cierta información puede o no ser sensible). Seguidamente se propone que el documento pase a un módulo "revisor" o de revisión por expertos, donde un experto del negocio tendría una interfaz adecuada para ver el documento tal como fue etiquetado con sus diferentes conceptos y datos sensibles identificados automáticamente. El experto revisor deberá allí aprobar el documento o rechazarlo, proveyendo en este último caso información que permita retroalimentar al sistema (alguna clasificación del tipo de error detectado). Finalmente se pasa el documento aprobado al módulo anonimizador, que cifra la información sensible con un algoritmo de clave pública o una función de hash. En la figura 1 se presenta un esquema gráfico de la arquitectura propuesta en este trabajo.

Módulo Etiquetador Posiblemente la mayoría de las arquitecturas para anonimizar documentos que se puedan pensar, debe contar con un componente que denominaremos "Etiquetador", que tiene la responsabilidad de detectar y delimitar los datos sensibles candidatos dentro de los documentos. Este módulo deberá utilizar alguna técnica para diferenciar a los datos sensibles de la información que no afecta la privacidad personas u organizaciones.

Para identificar los datos sensibles, se pueden pensar varias alternativas, con diferenciados grados de complejidad en su implementación:

1. Una primera aproximación sencilla (y poco efectiva), puede ser identificar nombres propios con la asistencia de diccionarios de términos. Se podrían identificar todas las palabras en un texto que no figuran en los diccionarios como candidatas a ser nombres propios o identificadores, y etiquetarlos de alguna forma. Un método así tendría un alto margen de falsos positivos, y dependería mucho de la calidad y tamaño de los diccionarios. Pero aún sería más óptimo que delimitar manualmente los datos sensibles

2. Una alternativa más costosa pero efectiva, consiste en identificar nombres propios o entidades con nombre ("Named Entities") utilizando clasificación automática de entidades (tecnologías muy vinculadas con el concepto de web semántica). En particular lo que se conoce con NER (Named Entity Recognition), es una técnica aplicable para delimitar nombres de personas, organizaciones o lugares geográficos, etc. Estas técnicas habitualmente se sustentan en procesamiento del lenguaje natural, técnicas basadas en gramáticas o modelos estadísticos.

Módulo Revisor Otro módulo que se puede identificar en una posible arquitectura de software para anonimización, es un componente al cual denominaremos "revisor". El objetivo de este módulo es presentar a un experto del negocio el resultado del proceso de etiquetado sobre un texto ya procesado, y permitir

aceptar o rechazar la inferencia de datos sensibles que pueda haberse hecho. Este módulo podría aprovechar el conocimiento del experto para retroalimentar al sistema, utilizando otro componente que presentaremos más adelante que llamaremos retroalimentador.

Módulo Anonimizador El módulo anonimizador lo que hace es concretamente sustituir los datos sensible del texto, previamente marcados y aceptados, por o bien algún término genérico, o por alguna referencia que permita revertir el proceso contando con una autorización o clave. En caso de que se realice lo último podríamos decir que el módulo sería un "despersonalizador" más que un anonimizador.

Retroalimentación La propuesta de ANONYMITEXT contempla la retroalimentación del sistema en base a la salida "rechazo" desde el módulo revisor. En caso de que un experto determine que el proceso de etiquetado no identificó correctamente algún término, debería existir una interfaz que le permita indicar el tipo de error que ocurrió y de esa forma reiniciar el proceso con esta nueva información incorporada.

4.2. Arquitectura MOSTAS

Se presentará otra propuesta de arquitectura que surge de un trabajo académico [13], vinculado a la identificación de términos biomédicos en documentos no estructurados en idioma español. Allí se trabaja sobre un framework de preprocesamiento de texto denominado MOSTAS, el cual se presenta también en otro artículo de interés [15]. A dicho framework se le adaptó un módulo de reconocimiento semántico de conceptos.

La arquitectura propuesta se ilustra en la figura 2.

En [15] el corrector ortográfico forma parte de un componente más grande junto con el anonimizador, y el analizador morfo-semántico se agrupa con otros componentes como se puede apreciar en el siguiente diagrama.

Más allá de las vistas y la evolución que evidentemente puede haber tenido el sistema MOSTAS desde 2008 a 2010 entre un artículo y el otro, se visualizan claramente algunos componentes en el sistema que se describirán a continuación.

Analizador morfo-semántico Este componente recibe las notas clínicas en formato no estructurado. Mediante una herramienta llamada STILUS, se realizan búsquedas de todas las palabras en un diccionario general del lenguaje español. De esta manera se identifican los términos generales que no tienen valor desde el punto de vista biomédico.

Motor de búsqueda de términos Las palabras no reconocidas por el analizador morfo-semántico, se buscan en otros diccionarios más específicos, de siglas, abreviaturas y acrónimos biomédicos. Si se encuentran definiciones en estos últimos, se guardan en un documento XML los significados. Para aquellas que no

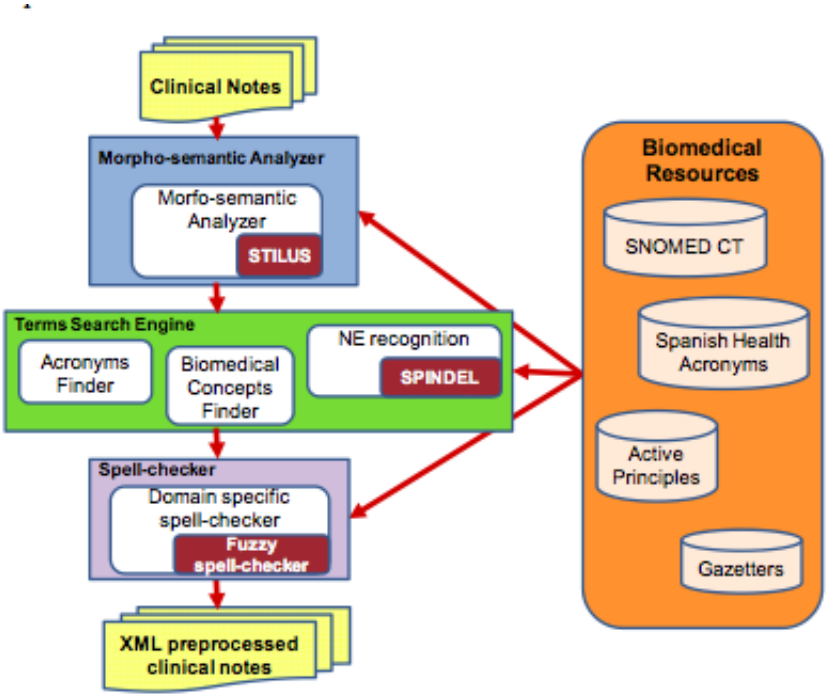

Figura 2: MOSTAS

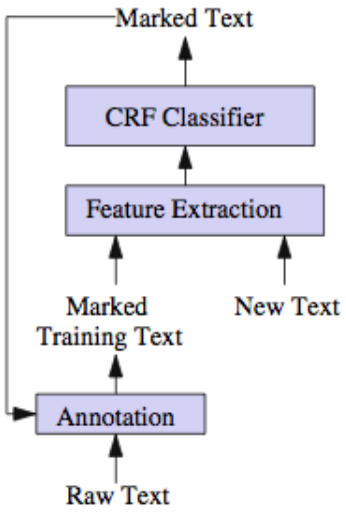

Figura 3: HIDE

se encontró una definición, se interopera con un sistema externo denominado "servidor de terminologías" (ST), con información de metatesauros como [23].

NER (Anonimización) Nuevamente, alimentado por los términos que aún no han sido reconocidos por el motor de búsqueda anterior, se procesan y se reconocen entidades nombradas (NER). En este componente se utiliza una vez más el software STILUS.

Verificador ortográfico El sistema MOSTAS añade un concepto interesante con este módulo corrector. Se plantea la posibilidad de que aquellos términos que no han sido reconocidos por los tres módulos anteriormente mencionados, pueden estar mal escritos. Lo que se hace es buscar por similitud los términos restantes en los recursos de información médica que se utilizaron anteriormente.

4.3. Arquitectura HIDE

Otra propuesta arquitectural la presenta el software de dominio público [HIDE] (Health Information DE-identification). Este framework opensource presenta herramientas para la anonimización y de-identificación de información médica no estructurada, y se puede descargar libremente bajo licencia MIT. La arquitectura propuesta se describe a alto nivel en [25].

Interfaz de anotación (Annotation) En primera instancia HIDE utiliza un proceso iterativo de etiquetado (tagging), basado en el entrenamiento mediante

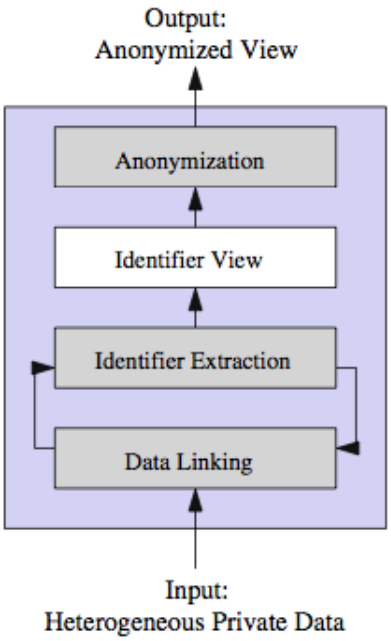

Figura 4: Clasificador HIDE

textos previamente etiquetados por un experto. Por lo tanto se ofrece un módulo que le permite al experto etiquetar textos identificando la información sensible.

Extracción de características (Feature Extraction) Este módulo lo que hace es identificar una serie de "características" (features) de interés para los términos que se encuentran, tanto en los textos etiquetados de entrenamiento como en los textos nuevos a ser anonimizados. Las características son por ejemplo el término en si mismo, la palabra anterior, la palabra siguiente, el uso de mayúsculas o no, si contiene caracteres especiales o no, si el término/token es un número, etc. Estas "features" luego serán utilizadas para clasificar el término.

Módulo Clasificador El texto con sus características, y los textos de entrenamiento alimentan un módulo que clasifica los términos, utilizando un método estadístico/probabilístico (CRF o Conditional Random Fields). La salida de este modulo es el texto etiquetado y listo para ser anonimizado. A partir de ese momento la arquitectura se refleja en el diagrama que se muestra en la figura 4.

Data Linking / Identifier Extraction La arquitectura de HIDE provee un módulo "Data Linker" que vincula toda la información sensible a una entidad. El sistema realiza un trabajo iterativo, extrayendo atributos del texto (con un

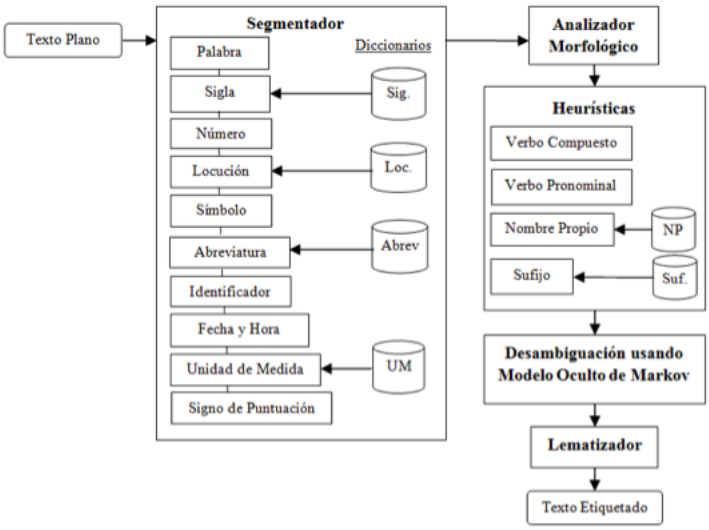

Figura 5: Etiquetador morfosintáctico para el español

módulo llamado Identifier Extractor), y vinculando los atributos a las entidades con el Data Linker. Lo que se hace es vincular internamente todos los atributos que son de una misma entidad (ejemplo, edad, dirección y nombre de una persona), generando las estructuras de datos adecuadas para representar a las distintas entidades nombradas que se encuentren en el texto. El módulo Linker utiliza técnicas probabilísticas para vincular los atributos, y se utiliza una base de datos externa con información de entidades para la tarea (ej. base de datos de personas, ubicaciones geográficas, etc).

Vista de Identificadores (Identifier View) Ya teniendo los atributos identificados y vinculados a entidades nombradas, se pasa a un módulo de presentación al usuario que se denomina "Vista de Identificadores". Allí se presenta el texto con sus identificadores y se permite elegir una estrategia de anonimización. Se permite realizar una anonimización completa o parcial (de algunos atributos críticos).

Anonimizador Finalmente el texto pasa al módulo anonimizador, el cual elimina los atributos sensibles en base al criterio indicado en la vista de identificadores.

4.4. Etiquetador morfosintático para el español

Si bien no se trata de un sistema anonimizador, la propuesta que realiza en su tesis de maestría la MSc. Lic. García Moya [29], describe una pieza de software que puede ser clave de un sistema de anonimización. Se propone un etiquetador morfológico, que básicamente lo que hace es etiquetar los términos en un texto libre, indicando si las mismas son, entre otras cosas, nombres propios. Teniendo identificado cada término según su tipo, la anonimización podría consistir en una etapa posterior que procese las etiquetas y sustituya algunas por información genérica no sensible. La propuesta arquitectural de García Moya para el etiquetador se describe en la figura 5.

Segmentador El módulo segmentador recibo como insumo de entrada un texto plano, y lo que hace es en primer lugar separar el texto en unidades lingüísticas tratables (letras, palabras, oraciones). También se realiza la primera clasificación de los términos, identificando por ejemplo si se trata de una palabra, un número, una fecha, un signo de puntuación, una sigla, etc, etc.

Analizador morfológico Este componente procesa cada palabra del texto, y le asigna una lista de posibles categorías gramaticales. Adicionalmente obtiene la información morfológica de la palabra (por ejemplo, género, número, persona, tiempo, modo, etc.). También se establece el lema de la palabra como un elemento más. Por lema se entiende a la forma "normal" que representa a una palabra, por ejemplo para un verbo su infinitivo, para un sustantivo su singular, para un adjetivo su masculino singular.

Heurísticas El proceso de etiquetado propuesto, incorpora el uso de heurísticas para determinar con mayor exactitud la naturaleza de las palabras procesadas por el analizador morfológico. Una heurística concreta que se propone, de gran interés para un proceso de anonimización, es la identificación de nombres propios utilizando un listado extenso de nombres y apellidos de personas, nombres de localidades, ciudades y países. Para la identificación del nombre propio se considera la aparición de la palabra en alguno de los listados mencionados, la presencia de una letra mayúscula al inicio de la palabra, la posición de la palabra dentro de la oración, el haber sido reconocida o no por el analizador morfológico, etc.

Desambiguador La desambiguación, un proceso bastante complejo desde el punto de vista de PLN, consiste en la determinación del significado de una palabra en un determinado contexto, cuando dicha palabra tiene múltiples significados. Una forma de identificar el significado, es utilizar un vasto conjunto de reglas (método deductivo), que en base a patrones permiten identificar el sentido de la palabra en su contexto. Estos métodos tienen un gran costo de desarrollo porque requieren definir un gran número de reglas para que sean efectivos. La alternativa consiste en utilizar métodos inductivos, que se basan en partir de

un conjunto de textos ("corpus") previamente etiquetados por un lingüista o experto, y en base a esto se entrena a un sistema de forma que luego permita aplicar el conocimiento sobre un texto nuevo original. En el caso particular de la propuesta de García Moya se opta por un método inductivo, utilizando el modelo estadístico "Modelo Oculto de Markov" (HMM por sus siglas en inglés) [PLA2001]. El método sin embargo es híbrido, tiene una componente deductiva, porque también se apoya en el uso de diccionarios.

Lematizador Al finalizar el proceso, éste módulo lematizador directamente asigna el lema que corresponda a cada palabra, proceso que es trivial cuando dicho lema fue determinado por el analizador morfológico. Para otros casos se siguen algunas reglas predefinidas. Por ejemplo para las siglas, se define como lema la el significado de la sigla (tomado de una base de datos definida para ello). Para las abreviaturas se devuelve la palabra expandida (ej, Dr. se lematiza como Doctor). Lo mismo ocurre con las unidades de medida. Finalmente si no fue posible etiquetar alguna palabra, se utiliza una etiqueta especial "[Desconocida]". Esta etiqueta especial podría según se sugiere alimentar algún componente corrector ortográfico, procedimiento visto también en la propuesta de arquitectura de MOSTAS.

4.5. Identificación de clusters de Named Entities

Cabe mencionar un tópico adicional, que no se presenta en forma específica en ninguna de las arquitecturas estudiadas. Se trata de la identificación de clusters para agrupar las Named Entities. Se trata de un problema que para la anonimización puede resultar muy relevante. Por formar clusters de Named Entities se entiende agrupar los términos que refieren a una misma Named Entity. Por ejemplo, en un texto podríamos encontrar a la misma persona referida por su nombre completo inicialmente, y posteriormente es habitual que ya se la refiera por uno de sus nombres o su primer apellido exclusivamente. Las organizaciones también pueden figurar de distintas formas a lo largo de un texto. Por ejemplo, Universidad de la República, o UDELAR, identifican a la misma Named Entity. Desde el punto de vista de la anonimización, si se sustituyen los N.E. por términos genéricos, resulta un valor agregado importante que se sustituyan todas las referencias a una misma identidad por un mismo término. Esto permite guardar una coherencia cuando se menciona a la entidad en el texto. De otra forma se puede realmente perder el sentido de lo expresado, es decir que la anonimización degradaría el valor del documento en el proceso. Este tema específico se estudia en un proyecto de grado de una universidad catalana [26].

Se podría pensar desde el punto de vista de la arquitectura de un sistema de anonimización, en un componente que se encargue de agrupar o formar clusters con las Named Entities luego de que las mismas son identificadas, de forma de conservar la coherencia en los textos anonimizados.

4.6. Conclusiones

Del análisis de las arquitecturas estudiadas, se recogen una serie de elementos en común, y otros específicos que se consideran de interés para ser incorporados en una posible arquitectura de referencia. Se analizarán estos aspectos a continuación.

4.6.1. Automatización de la anonimización

Como se ha visto en las distintas propuestas, existen numerosas técnicas tendientes a automatizar el proceso de anonimización de documentos.

Una clasificación de las mismas podría ser la siguiente:

1. Técnicas basadas en estadística y procesamiento de lenguaje.
2. Anonimización basada en reconocimiento de patrones, expresiones regulares y reglas.
3. Métodos basados en diccionarios.
4. Aprendizaje de máquinas / inteligencia artificial.
5. Combinaciones de las anteriores. Es frecuente la combinación de expresiones regulares y diccionarios.

Un elemento que llama la atención, es que se encuentran con facilidad numerosas publicaciones académicas relacionadas a la temática de la anonimización automática de documentos, pero en general trabajando sobre documentos en el idioma inglés. Son abundantes las referencias en relación al ámbito médico en Estados Unidos, en cuanto a los requerimientos impuestos por la ley HIPAA.

4.6.2. Aspectos comunes

Una de las primeras conclusiones que surgen, es que en la totalidad de las propuestas la piedra angular de la anonimización es un módulo o componente NER que aparece en todas las arquitecturas vistas, es decir la identificación de entidades con nombre. Este componente debe recibir como insumo un texto no estructurado, procesarlo, y brindar como salida el mismo texto pero con sus Named Entities identificadas mediante algún tipo de marca o etiqueta. De acuerdo a lo investigado, existen diversas herramientas que permiten realizar N.E.R., algunas con mayor o menor precisión en los resultados, y algunas proveen características adicionales como la clasificación de entidades con nombre. Cabe pensar entonces en una propuesta arquitectural que permita adaptar o intercambiar fácilmente estos "motores" N.E.R., es decir que podría pensarse en contar con alguna capa de abstracción, de forma de adaptar las interfaces a distintas herramientas y tecnologías.

En todas las propuestas específicas de arquitecturas de anonimización (ANONIMITEXT, MOSTAS y HIDE), se identifica un componente que realiza el procesamiento final del texto que es nada menos que el anonimizador. También se

Módulos	Anonimytext	MOSTAS	HIDE	Etiquetador ESP
NER	Si (Tagger)	Si (NE recognition)	Si (Ident. Extract)	Si (Analizador)
Corrector Ortográf.	No	Si (Spell-checker)	No	No
Heurísticas	Si (Laws)	Si (Terms S. Eng.)	No	Si (Heurísticas)
NE Clustering	No	No	No	No
Revisor	Si (Revision)	No	Si (Identifier View)	Si (Ident. View)
Anonimizador	Si (Anonymize)	Si (NE recognition)	Si (Anonymization)	No

Tabla 1: Aspectos comunes y específicos

visualizan especializaciones de este módulo, dado que la anonimización puede ser reversible o irreversible, parcial o total (en cuanto a los atributos que se anonimizan), de acuerdo a los requerimientos que se planteen cada caso.

4.6.3. Aspectos específicos

Luego del procesamiento de Named Entities, distintos enfoques arquitecturales proponen el concepto de retroalimentación del sistema de alguna forma. Uno de los enfoques que resultan interesantes en este sentido, propuesto por MOSTAS, es el de procesar los documentos utilizando algún corrector ortográfico cuando quedan elementos sin identificar. En la propuesta de [29], se aplican una serie de heurísticas luego de que el texto pasa por el analizador morfológico. Sería deseable en una arquitectura de referencia, que éstas posibles etapas de postprocesado del texto tuvieran interfaces de entrada y salida equivalentes, de forma que distintos componentes pudieran agregarse o quitarse tal como si fueran eslabones en una cadena.

Otro enfoque vinculado a la retroalimentación, en este caso asistida, es el de permitir a un experto identificar el tipo de errores cometidos en el proceso NER mediante alguna interfaz de usuario acorde, y luego retroalimentar el sistema con esta información.

En las distintas propuestas vistas, en algunos casos se utilizan analizadores morfológicos para identificar y clasificar las palabras, en otros casos herramientas basadas en diccionarios y motores de reglas, y también se vieron propuestas híbridas.

Finalmente, en relación al clustering de Named Entities, se puede pensar en un aspecto adicional a considerar en una arquitectura de referencia, que permita tener en cuenta este importante factor al anonimizar.

Se presentarán los aspectos comunes y específicos anteriormente mencionados en forma tabular en el cuadro 1.

5. Instanciación tecnológica de los módulos

Habiendo identificado los componentes tecnológicos comunes y específicos de las distintas arquitecturas estudiadas, en ésta sección se describen herramientas de software concretas que permiten instanciar los mismos. Se hace foco fundamentalmente en herramientas de uso libre que luego puedan ser utilizadas en un eventual prototipo de la arquitectura de referencia, pero se analizan algunas alternativas comerciales también.

5.1. TreeTagger

TreeTagger [22] es una herramienta de uso libre para investigación, educación y evaluación. Utiliza técnicas inductivas para etiquetar texto en numerosos idiomas luego de un proceso de entrenamiento. Se proveen archivos de parámetros con los resultados de entrenamiento para una gran cantidad de idiomas, entre ellos el castellano. El etiquetador permite identificar sustantivos propios, números, verbos, códigos alfanuméricos, y decenas de otras formas lingüísticas, así como identificar el lema de las palabras. Existen numerosos wrappers disponibles para integrar TreeTagger a software desarrollado en distintos lenguajes/plataformas como JAVA, Perl, Python, R, y Ruby.

Internamente TreeTagger utiliza técnicas basadas en árboles de decisión binarios para identificar las palabras [SCHMID1994].

Exista una interfaz gráfica (GUI) para Windows para interoperar fácilmente con TreeTagger, mediante la cual se configuran los numerosos parámetros que admite la herramienta.

5.2. FreeLing

FreeLing [30] es una suite de herramientas de análisis del lenguaje natural, open source y de uso libre bajo licencia GNU. También provee archivos de datos pre-entrenados para múltiples idiomas, entre ellos el español. Algunas de las funcionalidades que provee en particular para textos en idioma español:

1. Detección de oraciones.
2. Detección de números.
3. Identificación de fechas
4. Análisis morfológico
5. Detección de frases multi-palabra.
6. Detección de entidades nombradas (NE).
7. Clasificación de entidades nombradas (NE).
8. Etiquetado PoS (part-of-speech).

FreeLing se distribuye como una biblioteca C++, y se puede integrar fácilmente a software desarrollado en otros lenguajes como JAVA a través de la API JNI. Sin embargo también se distribuye una utilidad de línea de comandos para ejecutarlo como utilitario independiente.

5.3. STILUS

Stilus es una suite comercial de la empresa Daedalus, vinculada a la herramienta ANONIMYTEXT descrita en el artículo académico [41]. Stilus provee una gran cantidad de herramientas para el procesamiento del lenguaje natural, entre ellas Stilus NER para el reconocimiento automático de entidades con nombre (NER). Una característica diferente que provee Daedalus con Stilus, es que se proveen los productos en forma de servicios en la nube. Es decir que se pueden integrar los servicios de Stilus a software propio accediendo en línea a las herramientas, por ejemplo a Stilus NER. Cabe destacar que para el desarrollo de este trabajo se contactó a Daedalus y se proveyó por parte de la empresa de una clave para evaluar los productos con fines académicos.

5.4. Apache OpenNLP

OpenNLP [20] es un proyecto de la Apache Software Foundation, y por tanto naturalmente es de uso libre y open source, que consiste en una suite de herramientas para el procesamiento del lenguaje natural basadas en el aprendizaje de máquinas. Como la mayoría de las herramientas basadas en técnicas inductivas, se requiere entrenar los distintos componentes para luego utilizarlos sobre texto nuevo. Se pueden descargar sets de datos preentrenados para diversos idiomas, entre ellos el español.

Algunas de las herramientas que provee OpenNLP:

1. Detección de oraciones
2. Detección de entidades nombradas (NE)
3. Etiquetado PoS (part-of-speech)

5.5. OpenCalais

OpenCalais [42] es una API y un procesador semántico de documentos no estructurados. Es un producto propietario parte de la línea de productos ClearForest desarrollados por la Coporación Thomson Reuters, pero para éste servicio en particular se provee acceso libre al mismo tanto para uso personal como comercial. Provee una interfaz de web services mediante la cual es posible procesar documentos no estructurados (Texto / HTML / XML), y entre otras cosas, identificar y clasificar Named Entities. Identifica personas, organizaciones, ubicaciones geográficas, libros, álbumes musicales, autores, etc. La única limitante es que la API acepta hasta 50.000 transacciones por día para cada usuario registrado, y hasta 4 transacciones por segundo. En la figura 6 se presenta un

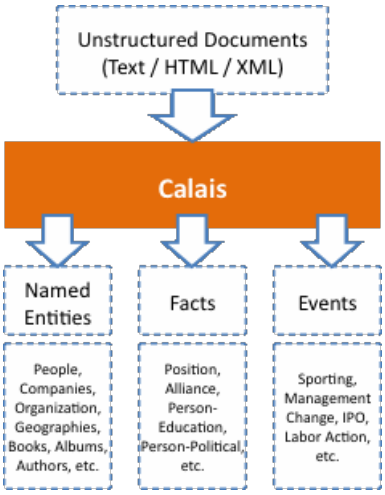

Figura 6: OpenCalais

esquema básico del funcionamiento del servicio:

Para poder probar el servicio sin necesidad de consumir el web service directamente desde una aplicación propia, se provee una utilidad llamada Calais Submission Tool que permite mediante su interfaz de usuario seleccionar documentos y enviarlos a Calais para su procesamiento. En pruebas realizadas con esta utilidad, se pudo observar una muy buena efectividad en la identificación de Named Entities. Coincide esta observación empírica con los resultados de un estudio de efectividad de sistemas NER, realizado en 2009 [34].

5.6. LingPipe

LingPipe[4] es un framework comercial desarrollado en JAVA puro. La empresa que lo desarrolla (alias-i) lo describe como "un kit de herramientas de procesamiento de texto usando computación lingüística".

Se presenta como una biblioteca integrable a cualquier aplicación, que provee una amplia gama de servicios, entre otros:

1. Tokenización
2. Etiquetado gramatical
3. Detección de Entidades con Nombre
4. Clustering
5. Identificación de frases importantes

6. Clasificación de tópicos

7. Minería de textos para bases de datos

8. Corrector ortográfico

Como se destaca, este framework provee dos características de interés para un sistema de anonimización. Tienen capacidad de identificar Entidades con Nombre, y además de agrupar texto en clusters.

LingPipe se puede utilizar de forma gratuita con fines académicos.

6. Propuestas para la documentación de la arquitectura

En el marco de éste trabajo se deberá documentar la arquitectura de referencia definida, por tanto corresponde seleccionar alguna de las propuestas existentes para dicha tarea. En principio se consideraron dos propuestas para la documentación de la arquitectura:

- Modelo de vistas "4+1" de Kruchten [39] : Se trata de una propuesta muy difundida para la documentación de una arquitectura, presente desde 1995. El modelo se centra en la especificación de casos de uso o escenarios de casos de uso, y la documentación de cuatro vistas orientadas a distintos stakeholders (Vista Lógica, de Desarrollo, de Procesos y Física). Resumidamente para describirlas, la Vista Lógica documenta el sistema en términos de los servicios que debe proveer para satisfacer los requerimientos funcionales. Habitualmente se representa con diagramas de clases y relaciones. La Vista de Procesos describe el sistema en términos de la dinámica de las tareas que se ejecutan, y el flujo dentro del mismo. La Vista de Desarrollo describe el sistema en términos de módulos, subsistemas, librerías o piezas de software, de forma que el mismo puede eventualmente ser distribuido para ser desarrollado por distintos desarrolladores, o equipos de desarrolladores. Finalmente, la Vista Física (o "de Despliegue") describe aquellos aspectos propios del sistema en su ambiente de ejecución, teniendo en cuenta elementos tales como servidores, bases de datos, el hardware en general que sea necesario, soluciones de alta disponibilidad y tolerancia a fallos, etc.

- Modelo de "Vistas y Perspectivas" de Rozanski / Woods [35]: La propuesta de Rozanski y Woods si bien comparte la idea de las Vistas (*Views*) de 4+1, añade un par de conceptos adicionales: los Puntos de Vista (*Viewpoints*) y las perspectivas (*Perspectives*). La propuesta consiste en describir la arquitectura a través de las Vistas (donde se puede definir un conjunto de ellas variable de acuerdo al proyecto, tomadas de un set más amplio de la propuesta 4+1). Los *Viewpoints* consisten en un conjunto de buenas prácticas, guías y plantillas que ayudan al arquitecto a definir las vistas seleccionadas. La propuesta añade un concepto novedoso, las Perspectivas, que suman una dimensión más a la descripción de la arquitectura. Mediante las Perspectivas, se logra describir aspectos inherentes a atributos de calidad del sistema que son transversales a distintos *Viewpoints*, y que deben ser considerados en etapas tempranas de la concepción del sistema. Ejemplos de Perspectivas propuestas por Rosanski y Woods son: Seguridad, Performance, Disponibilidad, Evolución, Accesibilidad, Localización, Regulación, etc.

7. Conclusiones finales

En las secciones precedentes se vieron diversas propuestas arquitecturales, donde fue posible determinar varios elementos comunes que se visualizan como componentes en una arquitectura de referencia. Vimos además que existen numerosas herramientas de software que permiten instanciar dichos componentes. Teniendo en cuenta todos estos elementos pasaremos ahora a definir nuestra arquitectura de referencia.

Como primera conclusión, cabe señalar que se identificaron una serie de requerimientos que se consideran drivers de la arquitectura. A continuación se describen dichos drivers.

7.1. Drivers de la Arquitectura

7.1.1. Adaptabilidad

Uno de los atributos de calidad que se visualiza como driver de la arquitectura es la adaptabilidad. Se pudo comprobar que el proceso de anonimización se apoya sobre el funcionamiento de algunos módulos o componentes que realizan determinados procesamientos específicos sobre un texto. Para cada uno de estos componentes se pudo ver también que existen muchas posibilidades en cuanto a su instanciación tecnológica. Es decir que se pueden utilizar distintas alternativas para cierto subproceso, optando por tal o cual producto o componente de software. Por tal motivo se entiende fundamental que la arquitectura permita adaptar con facilidad distintas opciones, aprovechando la diversidad y disponibilidad de productos que pueden desarrollar una misma tarea con distintos grados de calidad, costos o funcionalidades específicas.

7.1.2. Proceso

El proceso descrito anteriormente es determinante en la arquitectura. El subproceso clave consiste en la identificación de las Named Entities. Se parte de un texto no estructurado, y a partir de la aplicación de distintas técnicas se van determinando las N.E.. El proceso se retroalimenta constantemente, ya que cuando no se identifican ciertos términos se procesa el texto nuevamente mediante otros componentes (por ejemplo un corrector ortográfico, o reglas heurísticas), y se intentan determinar las N.E. sobre el resultado.

7.1.3. Configuración

La arquitectura debe representar un software configurable. Como se vio en secciones anteriores, el proceso de anonimización se puede ver como una secuencia o la colaboración de etapas de procesamiento del texto, en cada una de las cuales se identifican los datos sensibles utilizando distintas técnicas. Se debería poder configurar, acoplar y desacoplar estos componentes de acuerdo a su disponibilidad y las necesidades del usuario. También se debe poder confi-

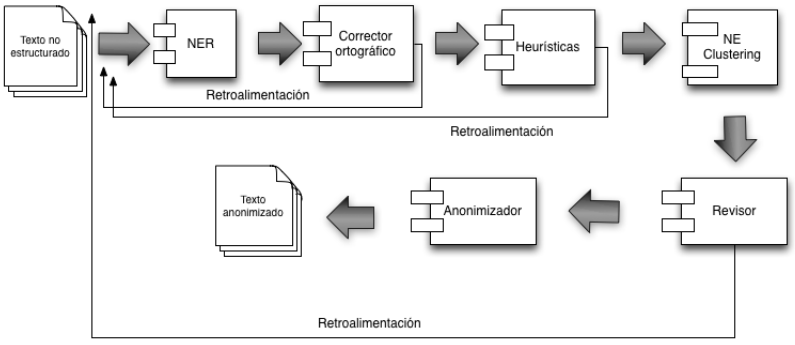

Figura 7: Diagrama de componentes comunes identificados

gurar el nivel de anonimización, el cual se especifica en el driver que se cita a continuación.

7.1.4. Diferentes Niveles de Anonimización

La arquitectura debe dar soporte a los distintos niveles de anonimización propuestos:

- Anonimización irreversible: Los datos sensibles son eliminados del documento, no es posible revertir el proceso de anonimización.
- Anonimización reversible: Se sustituyen los datos sensibles por referencias cifradas, permitiendo que entidades habilitadas puedan revertir el proceso u obtener las referencias a los datos sensibles.

A su vez, es posible considerar dos criterios de cobertura de la anonimización de acuerdo a los atributos que se incluyen en el proceso de anonimización:

- Anonimización total: Todos los datos identificatorios son anonimizados
- Anonimización parcial: Se anonimiza un subconjunto de los datos sensibles que surgen en el documento. El criterio de cuáles atributos se anonimizan debería ser configurable por ejemplo de acuerdo a lo que impongan distintas normas o reglamentaciones.

7.1.5. Diagrama de los principales componentes

Como segundo corolario de este estudio del estado del arte, en la figura 7 se presenta un diagrama donde se vinculan los principales componentes comunes que se identificaron, y aquellos específicos que se toman en consideración como de interés para una posible arquitectura de referencia.

Se visualiza en la propuesta presentada en ésta figura, que los componentes se encadenan unos a otros de forma secuencial, con varios puntos de retroalimentación donde se reinicia el proceso en alguna de las etapas anteriores.

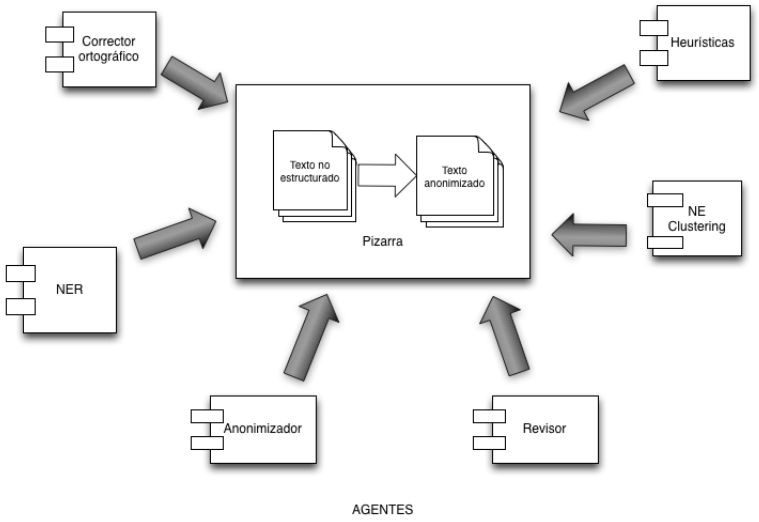

Figura 8: Diagrama de componentes vistos como un blackboard

La propuesta se asemeja bastante a la propuesta del patrón de diseño "Pipes and Filters".

Otra propuesta que se puede pensar es ver a los distintos componentes como “agentes” que trabajan en forma independiente sobre una fuente común (el texto), como se propone en el patrón de arquitectura “Blackboard” (Pizarra). En la figura 8 se ilustra la idea.

El patrón blackboard también se compone con una “estrategia”, que básicamente consiste en el método o forma en la cual se coordinan los distintos agentes para trabajar sobre la pizarra y la fuente de conocimiento (en éste caso el texto no estructurado).

7.2. Responsabilidades de los distintos componentes

Finalmente y como última conclusión, se describirán las responsabilidades que tienen los distintos componentes que se identificaron y que se integraron en los diagramas precedentes.

- **NER**: Este módulo identifica las Named Entities, es decir nombres propios, nombres y siglas referidas a organizaciones, ubicaciones geográficas, etc.
- **Corrector Ortográfico**: Un corrector ortográfico (spell-checker) tradicional, identifica errores ortográficos y los corrige automáticamente. La idea en este contexto de anonimización, es que de no poder clasificar algunas palabras del texto, se revise si las mismas no fueron mal escritas,

y se retroalimente al analizador. Este componente es opcional, ya que es una idea particular vista en tan solo una de las propuestas estudiadas. Se presenta como otra módulo que podría potencialmente ser útil en una implementación de un sistema de anonimización.

- **Heurísticas**: Este componente realiza una serie de procesamientos del texto para determinar patrones que pudieran inducir a errores al módulo NER. Una posibilidad es que se trate de un motor de "reglas" que al reconocer ciertos patrones en el texto pudiera identificar Named Entities no detectadas por el módulo NER, o por el contrario, determinar "falsos positivos" que pudiera haber identificado el NER.

- **NE Clustering**: La idea del módulo es agrupar las Named Entities en clusters, de forma que las distintas formas de referirse en el texto a la misma entidad estén vinculadas entre si, y no se identifiquen una misma N.E. como dos distintas.

- **Anonimizador**: Este componente sustituye las N.E. detectadas por nombres genéricos, o bien por el mismo dato pero cifrado (anonimización reversible).

- **Revisor**: Este componente presenta la posibilidad de realizar una revisión de la anonimización por parte de un experto, el cual podrá aprobar o rechazar el documento anonimizado.

Con este conjunto de insumos (drivers, componentes y responsabilidades), se está en condiciones de trabajar en la definición de la arquitectura de referencia para anonimización.

7.3. Modelo escogido para documentar la arquitectura

De los dos modelos estudiados, se escoge la propuesta de Rosanski y Woods para documentar la arquitectura de referencia en anonimización. Se considera que dicha propuesta se presenta como una visión más moderna e innovadora para documentar arquitecturas de software. Incorpora un conjunto de herramientas que permiten a un arquitecto de software expresar elementos medulares de la arquitectura, tales como los atributos de calidad. Adicionalmente se provee un conjunto de plantillas e información de referencia que simplifica la definición de la arquitectura, y se presenta como un modelo más flexible que no restringe las vistas a incluir.

Referencias

[1] Cavoukian A. and El Emam K. Dispelling the myths surrounding de-identification: Anonymization remains a strong tool for protecting privacy. Information & Privacy Commissioner of Ontario, June 2011. URL http://www.ipc.on.ca/images/Resources/anonymization.pdf.

[2] Grosskopf A., Decker G., and Weske M. *The Process: Business Process Modeling using BPMN*. Meghan Kiffer Press, 2009. URL http://www.bpmn-book.com.

[3] Sánchez A. Memetracker-web: Desarrollo de una interfaz web para un sistema de monotorización de la blogosfera política. URL http://e-archivo.uc3m.es/bitstream/10016/13797/1/MemoriaPFC_Memetracker_Web.pdf. Pág. 44.

[4] Alias-I. Lingpipe, 2003. URL http://alias-i.com/lingpipe/.

[5] BonitaSoft. Bonita open solution, 2001. URL http://es.bonitasoft.com/.

[6] Oracle Corporation. Mysql, 1995. URL http://dev.mysql.com.

[7] Cáceres D. Administración de bases de datos - tesis para optar el título de ingeniero civil, 2011. URL http://tesis.pucp.edu.pe/repositorio/bitstream/handle/123456789/943/NUNURA_CACERES_DIANA_ADMINISTRACION_BASE_DATOS.pdf?sequence=1.

[8] Reino de España. *Ley 14/2007, de 3 de julio, de investigación biomédica*. Colección Textos legales. Ministerio de Sanidad y Consumo, 2007. ISBN 9788476706886. URL http://books.google.com.uy/books?id=eP4xQwAACAAJ.

[9] Ministerio de Justicia y Derechos Humanos. Sistema argentino de informática jurídica, 2012. URL http://www.saij.jus.gov.ar/servicios/online/tesauro.htm.

[10] República Oriental del Uruguay. Ley nº 17.930 - presupuesto nacional 2005-2009, 2007. URL http://www0.parlamento.gub.uy/leyes/AccesoTextoLey.asp?Ley=17930&Anchor=.

[11] República Oriental del Uruguay. Ley 18.335 - pacientes y usuarios de los servicios de salud, Agosto 2008. URL http://200.40.229.134/leyes/AccesoTextoLey.asp?Ley=18335&Anchor=.

[12] República Oriental del Uruguay. Ley nº 18.331: Protección de datos personales y acción de "habeas data", 2008. URL http://www0.parlamento.gub.uy/leyes/AccesoTextoLey.asp?Ley=18331.

[13] Castro E., Iglesias A., Martínez P., and Castaño L. Automatic identification of biomedical concepts in spanish-language unstructured clinical texts. In *Proceedings of the 1st ACM International Health Informatics Symposium*. ACM, 2010. ISBN 978-1-4503-0030-8. doi: 10.1145/1882992.1883106. URL http://doi.acm.org/10.1145/1882992.1883106.

[14] Real Academia Española. Diccionario en línea de la real academia española., 2012. URL http://lema.rae.es/drae/?val=anonimizar.

[15] Iglesias A. et al. Mostas: Un etiquetador morfo-semántico, anonimizador y corrector de historiales clínicos. *Procesamiento de Lenguaje Natural*, 41(0), 2008. ISSN 1989-7553. URL http://sinai.ujaen.es/sepln/ojs/ojs/index.php/pln/article/view/2581.

[16] Troyano J.A. et al. Identificación de entidades con nombre basada en modelos de markov y árboles de decisión. *Procesamiento del lenguaje natural. Nº 31*, 2003. URL http://rua.ua.es/dspace/bitstream/10045/1552/1/PLN_31_28.pdf.

[17] Baladán F. Marco legal de la historia clínica electrónica en uruguay, 2011. URL http://www.agesic.gub.uy/innovaportal/file/1652/1/marco_legal_baladan.pdf.

[18] Pla F., Molina A., and Prieto N. Evaluación de un etiquetador morfosintáctico basado en bigramas especializados para el castellano. *Procesamiento de Lenguaje Natural*, (0), 2001. ISSN 1989-7553. URL http://sinai.ujaen.es/sepln/ojs/ojs/index.php/pln/article/view/3362.

[19] Sarasola F. Ley de protección de datos personales. *Universidad ORT Uruguay*, 2009. URL http://www.ort.edu.uy/fi/pdf/florenciasarasolalicsistemasort.pdf.

[20] Apache Software Foundation. Apache opennlp, 2000. URL http://opennlp.apache.org.

[21] Schmid H. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994. URL http://www.stttelkom.ac.id/staf/imd/Riset/POS%20Tagging/Using%20Decision%20Tree.pdf.

[22] Schmid H. Treetagger, 1994. URL http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

[23] IHTSDO. Snomed ct, 1999. URL http://www.ihtsdo.org/snomed-ct/.

[24] Div. Tecnología Informática. Base de jurisprudencia nacional pública, 12 2011. URL http://bjn.poderjudicial.gub.uy/.

[25] Gardner J. and Xiong Li. An integrated framework for de-identifying unstructured medical data. *Data Knowl. Eng.*, 68(12):1441–1451, December 2009. ISSN 0169-023X. doi: 10.1016/j.datak.2009.07.006. URL http://dx.doi.org/10.1016/j.datak.2009.07.006.

[26] Martínez Rodríguez J. Sistema de clustering de named entities. Master's thesis, Universidad Politécnica de Cataluña, 2008. URL http://www.recercat.cat/handle/2072/15414.

[27] González J.C. Anonimización: un enfoque útil para protección de la privacidad y de la confidencialidad, 2011. URL http://blog.daedalus.es/2011/06/21/anonimizacion-enfoque-util-proteccion-privacidad-confidencialidad/. DAEDALUS S.A. es la empresa comercial fundada por los creadores de la herramienta STILUS, utilizada por la arquitectura ANONIMYTEXT.

[28] Bass L., Clements P., and Kazman R. *Software Architecture in Practice*. Addison-Wesley, 2003. ISBN 9780321154958.

[29] García Moya L. Un etiquetador morfológico para el español de cuba. Master's thesis, Universidad de Oriente - Santiago de Cuba - Facultad de Matemática y Computación, 2008.

[30] Padró L. Freeling, 2003. URL http://nlp.lsi.upc.edu/freeling/.

[31] Rodríguez Yunta L. Bases de datos documentales: estructura y uso. *La información especializada en Internet - CINDOC/CSIC - Ministerio de Educación y Ciencia del Reino de España*, 2001.

[32] Poder Legislativo. Decreto 274/010, 2010. URL http://www.elderechodigital.com.uy/notas/ppla04.html.

[33] Xion Li and Gardner J. Hide (health information de-identification), 2008. URL http://code.google.com/p/hide-emory/wiki/Overview.

[34] Marrero M., Sánchez-Cuadrado S., Lara J., Morato, and Andreadakis G. Evaluation of named entity extraction systems. *Advances in Computational Linguistics. Research in Computing Science*, 41:47–58, 2009. URL http://site.cicling.org/2009/RCS-41/047-058.pdf.

[35] Rozanski N. and Woods E. *Software Systems Architecture: Working With Stakeholders Using Viewpoints and Perspectives*. Addison-Wesley Professional, 2005. ISBN 0321112296.

[36] United States of America. Health insurance portability and accountability act, 1996. URL http://www.gpo.gov/fdsys/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf.

[37] OMG. Business process model and notation (bpmn) versión 2.0, 2011. URL http://www.omg.org/spec/BPMN/2.0/PDF.

[38] Clements P., Kazman R., and Klein M. *Evaluating software architectures: methods and case studies*. SEI series in software engineering. Addison-Wesley, 2001. ISBN 9780201704822. URL http://books.google.com.uy/books?id=DV917BZ9RAgC.

[39] Kruchten P. Architectural blueprints - the "4+1"view model of software architecture. Paper published in IEEE Software, 11 1995. URL http://www.cs.ubc.ca/~gregor/teaching/papers/4+1view-architecture.pdf.

[40] Ruch P., Baud R., Rassinoux A., Bouillon P., and Robert G. Medical document anonymization with a semantic lexicon. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, January 2000. ISSN 1531-605X. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2244050\&tool=pmcentrez\&rendertype=abstract.

[41] Pérez-Laínez R, De Pablo-Sánchez C., and Iglesias A. *ANONIMYTEXT : Anonymization of unstructured documents*. KDIR 2009 - 1st International Conference on Knowledge Discovery and Information Retrieval, Proceedin, 2008.

[42] Thomson Reuters. Opencalais, 2008. URL http://www.opencalais.com.

[43] Meystre S., Friedlin F., , South B., Shen S., and Samore M. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, 2010. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2923159&tool=pmcentrez&rendertype=abstract.

[44] UNE/ISO. Une/iso 15489-1 información y documentación: Gestión de documentos, 2005. URL http://gestioninfo.wikispaces.com/file/view/UNE-ISO+15489-1.pdf.

[45] Muñoz Casals V. Sistemas de gestión documental, 2007. URL http://www.monografias.com/trabajos-pdf/sistema-gestion-documental/sistema-gestion-documental.pdf.