“Arquitectura de Referencia para Anonimizar Documentos”

Anexo B: Documento de Arquitectura de Software (S.A.D.)

Ing. Horacio Vico
horacio.vico@gmail.com
Versión 6.0 - Octubre de 2013.

Historial de Versiones

Versión	Fecha	Autor	Comentarios
1.0	Octubre de 2012	Horacio Vico	Versión inicial.
2.0	Mayo de 2013	Horacio Vico	Versión revisada.
3.0	Junio de 2013	Horacio Vico	Versión luego de revisión de D. Calegari.
4.0	Junio de 2013	Horacio Vico	Versión luego de procesada la segunda revisión de D. Calegari.
4.1	Junio de 2013	Horacio Vico	Se agregan parámetros de configuración adicionales del proceso en la Vista Funcional.
5.0	Julio de 2013	Horacio Vico	Versión anexo para documento principal.
6.0	Octubre de 2013	Horacio Vico	Versión luego de revisión de A. Delgado

Índice

Índice de figuras

Índice de tablas

1. Introducción

El incesante avance de las tecnologías de la información en el seno de las organizaciones, ha impulsado la incorporación de la Gestión Documental como una disciplina fundamental. El objetivo es optimizar la gestión y así maximizar el aprovechamiento de los grandes volúmenes de información que se encuentran en la forma de documentos. En algunos dominios de aplicación de la gestión documental tales como el gobierno electrónico, o los servicios de salud, entre otros, se presenta una necesidad recurrente: la anonimización. Este proceso consiste en proteger o incluso eliminar la información sensible contenida en los documentos.

La anonimización tiene aplicación en aquellos documentos donde la información de valor contenida en ellos, es independiente de los datos personales o la información sensible. El fin es que dicha información pueda ser utilizada dentro de la propia organización o por terceros, sin que esto implique vulnerar la privacidad y la confidencialidad de los datos personales de las personas físicas o jurídicas que se referencian en el documento original. Algunos países poseen legislación muy específica vinculada con la anonimización. En Uruguay se ha aprobado normativa referente a la protección de datos personales, exigiendo a las organizaciones garantizar la confidencialidad de los datos personales que manejan. Este tipo de normas jurídicas han impulsado la investigación y el desarrollo de técnicas y metodologías para la anonimización automática o semiautomática de los documentos.

El problema informático de anonimizar documentos no resulta trivial, más teniendo en cuenta que muchos de ellos no siguen un formato estructurado que permita identificar fácilmente la información sensible dentro de los mismos. Disciplinas computacionales tales como el procesamiento de lenguaje natural, la minería de textos, o el aprendizaje automático por máquinas, se presentan como herramientas aplicables para la resolución de este tipo de problemas. Desde el punto de vista de la arquitectura de software, la integración de diferentes elementos tecnológicos que se pueden utilizar en un proceso de anonimización tales como los mencionados, representa un tema de investigación en sí mismo.

En el marco de este proyecto, fueron estudiadas diversas propuestas de arquitecturas de anonimización tales como ANONIMYTEXT[11], MOSTAS [2], HIDE [9], y Etiquetador ESP[8]. De dichas propuestas se identificaron características comunes de los sistemas de anonimización, y se seleccionaron aquellas que se consideran de utilidad para la definición de una arquitectura de referencia, complementándolas con definiciones específicas de la propuesta que aquí se describe.

1.1. Objetivo y Alcance

Esta propuesta surge en el contexto de un trabajo académico de tesis, cuyos objetivos específicos abarcan el estudio del problema de la anonimización en el marco de la gestión documental, el análisis de arquitecturas de anonimización existentes, y la descripción de una arquitectura de referencia que recoja las mejores ideas vistas, con los aportes originales que se pudieran proponer.

El presente documento describe dicha arquitectura de referencia para un sistema para anonimizar documentos.

El alcance de la propuesta abarca la descripción de la arquitectura en forma genérica. No se presenta un sistema en particular sino la arquitectura en la cual podrán basarse implementaciones concretas. Por tal motivo se seleccionará un conjunto de vistas tendientes a describir aspectos generales de la arquitectura, y no aquellos específicos que serán de particular interés a la hora de implementar un sistema de anonimización concreto.

Durante el estudio del estado del arte en anonimización, se determinaron una serie de drivers que condicionan la arquitectura de referencia que se propone.

Se resumirán dichos drivers a continuación:

1. **Adaptabilidad**: En distintas propuestas de anonimización estudiadas, se pudo visualizar al proceso de anonimización como una cadena de subprocesos donde se aplican distintas herramientas sobre el texto del documento. A su vez, se estudiaron distintas instanciaciones tecnológicas para cada uno de esos subprocesos, y se encontró gran variedad de herramientas que permiten realizar cada una de estas tareas con distintas ventajas y desventajas (términos de licenciamiento, efectividad, nivel de configuración, etc). Una arquitectura de referencia podría beneficiarse mucho si permitiera abstraer de alguna manera la interacción (interfaces) entre estos subprocesos, de forma que se pudiera optar por uno u otro en base a configuración.

2. **Proceso**: El proceso es otro conductor de la arquitectura. En el contexto del proyecto fueron estudiadas diversas propuestas de arquitecturas para sistemas de anonimización, donde se pudo ver un común denominador: un proceso en etapas, donde el texto va pasando de un módulo hacia otro hasta que se logra determinar las entidades con nombre (*Named Entities*), y las mismas pueden finalmente ser anonimizadas. No se encuentran elementos que permitan pensar en una propuesta diferente para una arquitectura de referencia, por tanto la misma deberá soportar un proceso similar al visto en las arquitecturas preexistentes. El proceso descrito será detallado en la vista funcional más adelante en éste documento.

3. **Configuración**: Del driver anterior se desprende otro aspecto conductor de la arquitectura. Se busca una propuesta de arquitectura que permita optar por configuración por diferentes herramientas que se puedan utilizar para realizar los distintos procesamientos que se realizan sobre el texto. Además es posible que incluso alguna de éstas etapas pudiera directamente activarse o desactivarse, de acuerdo a los requerimientos del usuario en cada caso.

Como se podía inferir de la descripción del driver “Proceso”, la arquitectura que se propone se basa en la idea propuesta por el patrón de diseño “Pipes and

Filters" o "Pipeline", presentando una cadena de subprocesos cuya entrada es la salida de otro subproceso. Esta idea coincide con la propuesta de todas las arquitecturas estudiadas.

1.2. Propuesta de diseño de la arquitectura

Para el diseño y la descripción de ésta arquitectura, se sigue la propuesta de Rozanski y Woods [12], documentando los puntos de vista (ViewPoints) que se determinaron son de interés para reflejar una arquitectura genérica como la propuesta, y teniendo en cuenta los intereses de los stakeholders:

1. **Vista Funcional**: Mediante esta vista se describirá la responsabilidad de cada componente, sus interfaces e interacciones con el resto del sistema, y fundamentalmente sus responsabilidades.

2. **Vista de Información**: Describirá la forma en que se fluye la información en el sistema. En el caso de la arquitectura existirá un flujo constante de información proveniente del texto que se está procesando. Se entiende que mediante ésta vista se podrá describir el mismo en detalle, incluyendo el formato y estructura de la información, su flujo dentro del sistema, etc.

3. **Vista de Desarrollo**: Describirá aspectos de interés para los interesados (stakeholders) involucrados en la implementación o instanciación de la arquitectura propuesta, tales como la organización de los módulos y su estructura.

2. Glosario

Término	**Definición**
Stakeholder	En el contexto de la arquitectura de software, persona o entidad que tiene un interés sobre el sistema cuya arquitectura se describe (ejemplo: usuario, cliente, sponsor, etc)
Anonimización	Proceso por el cual deja de ser posible establecer por medios razonables el nexo entre un dato y el sujeto al que se refiere.[3]
Blackboard	(Pizarra), patrón de arquitectura de utilidad para resolver problemas donde no se conoce una solución determinística. Varios subsistemas especializados (agentes) se conjugan para construir una solución aproximada o parcial. [5]
NER	Named Entity Recognition, reconocimiento de entidades con nombre. Proceso de identificación de las entidades con nombre en un texto.
Entidad con Nombre	Nombres de personas, organizaciones, localizaciones, expresiones de horas, cantidades, valores monetarios, porcentajes, etc.
Pipes and Filters	Patrón de diseño, donde un proceso se subdivide en una secuencia de subprocesos independientes (filtros), los cuales se conectan a través de canales (pipes).

3. Contexto del Sistema

3.1. Entorno del Sistema

En la Figura 1 se presenta un diagrama donde se pueden ver los "actores" típicos que interactúan con un sistema de anonimización. El software de anonimización presenta interfaces hacia otros sistemas, típicamente sistemas de gestión documental, mediante las cuales recibe la información (documentos) a anonimizar, y luego de procesarlos los devuelve al sistema de origen. No se planteo como un objetivo de un sistema de anonimización almacenar la información anonimizada ni presentarla a usuarios finales, la idea es que la anonimización sea un servicio semiautomático que puede ser utilizado por otros sistemas de la organización. Como único actor humano del sistema de anonimización, se presenta un experto del dominio, que en caso de que el sistema de anonimización contemple una etapa de validación será quien deberá aprobar o rechazar el texto procesado por el sistema, y retroalimentar si así corresponde al propio software.

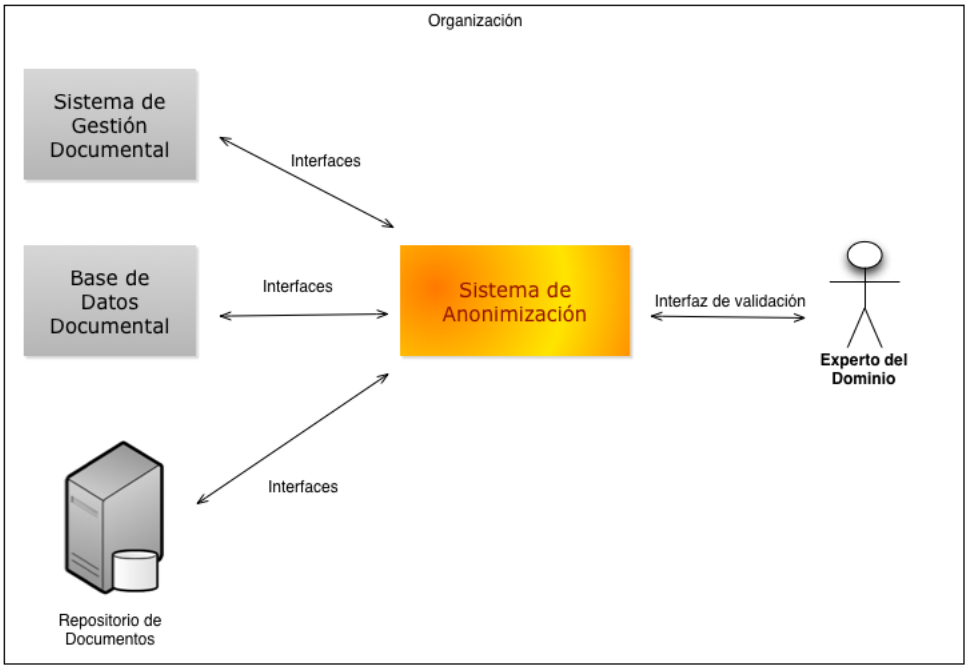

Figura 1: Diagrama de Contexto

3.1.1. Stakeholders (interesados)

En esta sección se enumeran los stakeholders que se visualiza tendrán interés en el sistema descrito, y hacia quienes se orienta la información contenida en el

presente documento.

- **Arquitectos de Software**: Por tratarse de una arquitectura genérica "de referencia", los principales interesados en la presente arquitectura son arquitectos de software que pretendan diseñar sistemas para la anonimización de documentos de texto.
- **Desarrolladores:** Programadores o analistas que fueran a implementar alguna "instanciación" de ésta arquitectura de referencia en un sistema concreto. Puede ser de interés conocer los conceptos de alto nivel y los drivers que llevaron a definir la estructura general de la arquitectura.
- **Estudiantes o personal del ámbito académico:** Interesados en la temática de la anonimización o las arquitecturas de software, así como lógicamente los directamente interesados en el presente trabajo de tesis (tutor, tribunal, etc.)

3.2. Visión General de los Requerimientos

Se enumeran los requerimientos funcionales y no funcionales identificados para un sistema de anonimización.

Referencia	Descripción del Requerimiento
RF1	El sistema debe poder procesar documentos no estructurados, en formato texto plano.
RF2	Se deben poder utilizar distintas herramientas de software para el procesamiento del texto indistintamente. La herramienta a utilizar en cada etapa debe ser un elemento más de configuración.
RF3	El sistema debe permitir a un experto del dominio validar la salida anonimizada, y permitir aprobar el documento o reprobarlo brindando feedback que retroalimente al sistema.
RF4	Se debe poder configurar el alcance de la anonimización, la cual puede ser total (todos los atributos que identifican a personas u organizaciones), o parcial (un subconjunto de los atributos).
RF5	El sistema debe permitir por configuración, realizar una anonimización irreversible (se elimina por completo la información sensible), o reversible (se sustituye la información sensible por una referencia cifrada a la misma).
RF6	El sistema debe permitir la introducción de reglas o patrones (heurísticas) definidas por un usuario experto, para contemplar características específicas de los documentos de un dominio en particular.
RF7	El sistema debe contemplar la posibilidad de interactuar con fuentes externas, mediante adaptadores, que provean información de términos específicos de un dominio (diccionarios, tesauros, gacetillas, acrónimos, etc.)
RNF1	Adaptabilidad: Para dar soporte al RF2.
RNF2	Configurabilidad: Para dar soporte a RF2 y RF5
RNF3	Interoperabilidad: El sistema deberá interoperar con distintas fuentes externas (RF7)
RNF4	Documentación: Las interfaces del sistema deberán estar debidamente documentadas, de forma que un usuario avanzado pueda agregar nuevas fuentes de conocimiento o herramientas de procesamiento de lenguaje natural.
RNF5	Extensibilidad: El añadir nuevas herramientas de procesamiento de lenguaje (RF2 y RNF1) deberá ser un proceso sencillo, que no implique modificaciones mayores al sistema.
RNF6	Auditoría: El sistema deberá registrar las interacciones de usuarios con los documentos.
RNF7	Seguridad: Los documentos no anonimizados no deben ser accedidos por usuarios no autorizados.

3.3. Escenarios del Sistema

En esta sección se listarán algunos escenarios funcionales y no funcionales, describiendo situaciones que el sistema debe afrontar y la respuesta esperada ante la misma.

3.3.1. Escenarios Funcionales

Referencia	ESRF1
Requerimiento vinculado	RF1
Visión General	Cómo el sistema procesa documentos no estructurados, en formato texto plano.
Estado del Sistema	Normal
Entorno del Sistema	El entorno del sistema opera normalmente.
Estímulo externo	El sistema recibe un documento no estructurado en texto plano.
Respuesta	El sistema debe ser capaz de procesar el documento, e identificar las Entidades con Nombre contenidas en el mismo, para su posterior anonimización.

Referencia	ESRF2
Requerimiento vinculado	RF2
Visión General	Cómo el sistema se adapta para contemplar distintas herramientas de software de PLN.
Estado del Sistema	Normal
Entorno del Sistema	El entorno del sistema opera normalmente.
Estímulo externo	El usuario desea intercambiar el etiquetador de TreeTagger por el de OpenCalais
Respuesta	El usuario accede a una interfaz de configuración del sistema y selecciona la herramienta de etiquetado

Referencia	ESRF3
Requerimiento vinculado	RF3
Visión General	Cómo el sistema permite a un usuario experto validar la salida anonimizada.
Estado del Sistema	Normal
Entorno del Sistema	El entorno del sistema opera normalmente.
Estímulo externo	El usuario experto opera el sistema, y visualiza un texto anonimizado por el mismo. Identifica errores en la identificación de un término como Entidad con Nombre.
Respuesta	El usuario experto marca el término como incorrectamente identificado, y retorna el documento para nuevo procesamiento. El sistema se retroalimenta con la información provista por el experto, y no vuelve a identificar el término como Entidad con Nombre.

Referencia	ESRF4
Requerimiento vinculado	RF4

Visión General	Cómo el sistema permite configurar el alcance de la anonimización.
Estado del Sistema	Normal
Entorno del Sistema	El entorno del sistema opera normalmente.
Estímulo externo	El sistema está configurado para anonimizar todos los datos que identifica como sensible. El usuario del sistema desea configurar que los números que figuren en el texto no sean anonimizados.
Respuesta	El usuario accede a una interfaz de configuración del sistema y define que los números no serán anonimizados.

Referencia	ESRF5
Requerimiento vinculado	RF5

Visión General	Cómo el sistema permite configurar distintos niveles de anonimización.
Estado del Sistema	Normal
Entorno del Sistema	El entorno del sistema opera normalmente.
Estímulo externo	El sistema anonimiza de forma irreversible y total. Se desea optar por anonimización reversible, contemplando un subconjunto de datos sensibles de acuerdo a cierta legislación o norma.
Respuesta	Un administrador del sistema configura el proceso de anonimización como reversible, y enumera los atributos que serán anonimizados de acuerdo a una serie de criterios (nombres propios, información geográfica, números, etc).

Referencia	ESRF6
Requerimiento vinculado	RF6
Visión General	Cómo el sistema se permite al usuario introducir patrones, heurísticas o reglas para contemplar casuísticas que surge en un dominio particular.
Estado del Sistema	Normal
Entorno del Sistema	El entorno del sistema opera normalmente.
Estímulo externo	El experto de dominio en la interfaz de revisión detecta que fue marcada como información a anonimizar el nombre Gabriel García Márquez, autor de un texto al cual se hace referencia en el documento.
Respuesta	El usuario accede a una interfaz de configuración del sistema, define una nueva regla por la cual el nombre Gabriel García Márquez no será considerado como nombre a anonimizar.

Referencia	ESRF7
Requerimiento vinculado	RF7
Visión General	Cómo el sistema puede utilizar fuentes externas de conocimiento para mejorar la identificación de Entidades con Nombre.
Estado del Sistema	Normal
Entorno del Sistema	El entorno del sistema opera normalmente.
Estímulo externo	El sistema será utilizado en un contexto de ciencias biomédicas. Se desea utilizar la fuente SNOMED de terminología clínica en inglés.
Respuesta	Un programador desarrolla un componente adaptador que implemente una interfaz documentada del sistema de anonimización. Se configura la utilización de esta nueva interfaz.

3.3.2. Escenarios no funcionales

Referencia	ESRNF4
Requerimiento vinculado	RNF4
Visión General	Cómo el sistema posee documentación adecuada para sus interfaces que permite extenderlo con facilidad.
Estado del Sistema	Normal
Entorno del Sistema	El entorno del sistema opera normalmente.
Estímulo externo	Un desarrollador desea programar una funcionalidad que permita al sistema interoperar con una fuente de conocimiento nueva (ejemplo, un diccionario o tesauro en línea).
Respuesta	El desarrollador estudia la documentación del sistema, y desarrolla un módulo que respete las interfaces y estructuras de datos definidas. El sistema interopera con la nueva fuente de conocimiento.

Referencia	ESRNF5
Requerimiento vinculado	RNF5
Visión General	Cómo el sistema permite adaptar una nueva herramienta de software de procesamiento de texto, sin necesidad de realizar cambios en el diseño.
Estado del Sistema	Normal
Entorno del Sistema	El entorno del sistema opera normalmente.
Estímulo externo	Se presenta una nuevo software de reconocimiento de entidades con nombre, el cual se desea pueda ser utilizado en el proceso de anonimización del sistema.
Respuesta	Un desarrollador estudia la documentación del sistema, y desarrolla un módulo adaptador que respete las interfaces adecuadas. La herramienta se añade como una opción más en la lista de herramientas de reconocimiento de entidades por nombre, y el usuario puede optar por ella en la configuración del sistema.

Referencia	ESRNF7
Requerimiento vinculado	RNF7
Visión General	Cómo el sistema audita las interacciones de los usuarios con los documentos.
Estado del Sistema	Normal
Entorno del Sistema	El entorno del sistema opera normalmente.
Estímulo externo	Un usuario experto, en la interfaz de validación del sistema, identifica errores en la anonimización y retroalimenta el sistema.
Respuesta	El sistema registra en su log de auditoría el nombre del usuario que realizó la operación, y guarda los cambios que se realizaron al documento.

Referencia	ESRNF8
Requerimiento vinculado	RNF8
Visión General	Cómo el sistema permite proteger la información confidencial contenida en los documentos que procesa.
Estado del Sistema	Normal
Entorno del Sistema	El entorno del sistema opera normalmente.
Estímulo externo	Un usuario no autorizado intenta ver los documentos originales sin anonimizar.
Respuesta	El sistema rechaza el acceso, y el incidente es registrado en la auditoría.

4. Condicionantes de la Arquitectura

4.1. Metas

El proceso de anonimización implica analizar un texto identificando cada referencia a persona, lugar u organización cuyos datos se desean proteger, sustituyéndolos por una versión cifrada de los mismos (cifrado reversible), o directamente por nombres genéricos que no permitan identificarlos. El sistema de anonimización tiene como meta principal implementar un mecanismo automático o semiautomático, que permita eliminar la información sensible en documentos no estructurados.

- Como *business driver* se presenta el avance de legislación específica en lo que refiere a la protección de datos personales. De acuerdo al estudio del estado del arte realizado, se verificó la existencia en otros países de normas específicas orientadas a la anonimización de información de pacientes en historias clínicas. En Uruguay se ha comenzado a legislar en la materia[4], y si bien aún la normativa es algo genérica, es de esperarse que a medida que se vayan reglamentando las leyes surjan requerimientos específicos que puedan ser atendidos por procesos de anonimización.

- La definición de una arquitectura de referencia además, pretende tomar las mejores ideas de variadas arquitecturas estudiadas, y proponer un diseño de alto nivel que por sobre todo sea adaptable e interoperable con distintas herramientas de software preexistentes, y con la documentación necesaria y un diseño que soporte que pueda ser adaptado a futuras herramientas sin un que implique un rediseño del sistema.

4.2. Principios arquitecturales

Complementando la definición de requerimientos presentada en la sección anterior, se enumeran a continuación algunos principios que condicionan la definición de la arquitectura. Los principios proponen consideraciones generales que tienen un alcance global en la arquitectura. Más allá de los drivers y requerimientos no funcionales identificados previamente, estos principios definen pautas que afectan globalmente al sistema, que se deberían tener en cuenta en todo momento al implementar el sistema descrito por esta arquitectura.

Referencia	Principio 1 - Interoperabilidad
Principio	El software a desarrollar debe seguir estándares que promuevan la interoperabilidad de las aplicaciones y los datos.
Razón	El sistema deberá interoperar con diversas herramientas de software externas. El apego a estándares simplifica la interoperabilidad.
Implicancias	▪ Se utilizarán estándares de la industria para el intercambio de datos, siempre que sea posible.

Referencia	Principio 2 - Usabilidad
Principio	La interfaz de usuario debe ser sencilla y clara. La tecnología que se utilice debe ser transparente al usuario del sistema.
Razón	Un sistema con una interfaz de usuario sencilla mejora la productividad del usuario, y reduce los tiempos de aprendizaje de nuevos usuarios.
Implicancias	▪ La interfaz de usuario del sistema debe ser uniforme, respetar un mismo “look&feel”. ▪ La complejidad de las herramientas con las que se interopere debe ser encapsulada por el sistema.

Referencia	Principio 3 - Independencia tecnológica
Principio	El sistema descrito deberá poder implementarse en diferentes plataformas de desarrollo.
Razón	Independizar el sistema de una plataforma específica, aporta el beneficio directo de que pueda ser utilizado en un mayor número de escenarios. Por tratarse de una arquitectura genérica debe describir un sistema que luego pueda implementarse sobre plataformas específicas, y no condicionar dicha implementación. Este principio además está vinculado al principio de Interoperabilidad definido previamente.
Implicancias	▪ Se deberá evitar la integración de herramientas que solo se puedan utilizar en una plataforma específica. ▪ Se deberán excluir definiciones que puedan condicionar a la utilización de una plataforma específica.

5. Vistas de la Arquitectura

Presentados los principales requerimientos funcionales, atributos de calidad, y los principios que rigen esta arquitectura, se describirá el sistema de anonimización utilizando el subconjunto de vistas que se entienden son relevantes, teniendo en cuenta los stakeholders identificados y las principales características del sistema de anonimización.

5.1. Vista Funcional

En primer lugar se presenta la Vista Funcional de la arquitectura. A través de la misma, se introducen los diferentes elementos funcionales del sistema de anonimización, sus responsabilidades y la forma en que se comunican entre sí. Asimismo se presentan los parámetros de configuración utilizados por los componentes funcionales.

5.1.1. Componentes Funcionales del Proceso de Anonimización de Documentos

El proceso fue definido como uno de los principales drivers de la arquitectura. En el contexto del proyecto fueron estudiadas diversas propuestas de arquitecturas para sistemas de anonimización, donde se pudo identificar un común denominador: un proceso en etapas, donde el texto va pasando de un módulo hacia otro hasta que se logra reconocer las entidades con nombre, y las mismas pueden finalmente ser anonimizadas.

En primer lugar el sistema de anonimización deberá tener algún tipo de interfaz de usuario (UI), donde presentará los documentos en sus estados iniciales y anonimizados. Los documentos podrían ser ingresados directamente por los usuarios de alguna manera, o provendrían de fuentes externas de documentos, tales como un sistema de gestión documental o una base de datos documental.

La interfaz de usuario deberá permitir además parametrizar y configurar al sistema de acuerdo a las necesidades específicas del usuario. Esta interfaz de usuario además deberá permitir manejar distintos niveles de seguridad, es decir que se deberá proveer alguna interfaz de autenticación de usuarios, para contemplar la diferenciación entre los usuarios comunes y los expertos/validadores que se mencionaron en la especificación de requerimientos.

Por otra parte, la lógica de negocios y el motor del sistema encargado de gestionar el proceso, se puede abstraer en una capa del sistema donde encontraremos diversos módulos o componentes que se describirán a continuación.

El núcleo del proceso, y por ende de esta capa lógica del sistema, se centra en el reconocimiento de las entidades con nombre, que se puede definir como "una subtarea de la recuperación de información cuyo objetivo es localizar y clasificar los elementos atómicos en el texto sobre categorías predefinidas como nombres de personas, organizaciones, localizaciones, expresiones de horas, cantidades, valores monetarios, porcentajes, etc."[1]

El componente que llamaremos NER, tiene la responsabilidad de procesar el reconocimiento primario de entidades con nombre, y debe recibir como insumo un texto no estructurado, procesarlo, y brindar como salida el mismo texto pero con sus entidades con nombre identificadas mediante algún tipo de marca o etiqueta. De acuerdo a lo investigado, existen diversas herramientas de procesamiento del lenguaje natural, que permiten realizar NER, algunas con mayor o menor precisión en los resultados. Estas herramientas utilizan técnicas basadas en estadísticas, aprendizaje de máquinas, inteligencia artificial, diccionarios o combinaciones de dichas técnicas, para identificar las entidades con nombre en el texto.

Algunas herramientas proveen características adicionales como la clasificación de entidades con nombre. La clasificación no es otra cosa que categorizar las entidades con nombre de acuerdo a su naturaleza (por ejemplo determinar si la entidad con nombre se corresponde a un nombre propio, una localización geográfica, un identificador único de una persona, etc.).

Se pensó entonces en un proceso que permita adaptar o intercambiar fácilmente estos "motores" NER, contando a su vez con una capa de abstracción, de forma de adaptar las interfaces a distintas herramientas y tecnologías.

El proceso contempla además la integración de herramientas que permitan agrupar las entidades con nombre en clusters (módulo "Clustering"). Esto permite por ejemplo que conceptos equivalentes (por ejemplo una sigla y su significado, "O.N.U. = Organización de las Naciones Unidas"), puedan ser identificados como una misma entidad con nombre, y en consecuencia la anonimización aplique el mismo criterio para las equivalencias, aportando a conservar de mejor manera la coherencia del documento anonimizado.

También se incorpora la identificación de entidades con nombre adicionales o específicas mediante el uso de heurísticas, tales como la identificación de reglas, expresiones regulares o patrones específicos, que puedan ser personalizados de acuerdo al dominio del cuerpo de texto a analizar. En el proceso se definen tareas específicas en este sentido, y se deja la puerta para integrar herramientas adicionales que se pudieran adaptar para determinado dominio específico (diccionarios o tesauros, categorización mediante servicios web, etc).

En todas las propuestas específicas de arquitecturas de anonimización, se identifica otro componente mediante el cual se realiza el procesamiento final del texto, que es concretamente el módulo anonimizador. Se 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.

Como se mencionó anteriormente, la lógica del sistema deberá interoperar con diversas fuentes y herramientas externas, tales como las herramientas NER o las herramientas de clustering.

Toda la interoperabilidad del sistema, ya sea con herramientas externas, fuentes de conocimiento, o las propias fuentes de documentos, se gestionará con una capa de integración que denominaremos "Conectores".

Finalmente, los requerimientos del sistema especifican la necesidad de contar con auditoría en el proceso, así como seguridad para los documentos que se

procesan. Estos dos componentes son afectan al sistema en distintos niveles, y los podemos visualizar como aspectos o capas transversales.

Vinculado a la seguridad, en el sistema se deberían definir los siguientes roles, de acuerdo a los requerimientos RF3 y RNF7:

1. Rol "Usuario": Rol genérico a asignar a los usuarios del sistema. Permite procesar documentos en el sistema de anonimización con la configuración predeterminada.

2. Rol "Revisor": Este rol se asigna a los expertos del dominio de los documentos que se procesan. Tiene la potestad de validar un documento procesado por el sistema de anonimización. El usuario revisor podrá además retroalimentar al sistema de manera de que se enriquezca el proceso de anonimización con su conocimiento, y además podrá reconfigurar el sistema de acuerdo a las necesidades.

En base a todos los elementos mencionados, en la Figura 2 se ilustra el sistema y los componentes descritos en base a un modelo de capas clásico. Allí se puede visualizar:

1. La interfaz de usuario gestionada por la capa UI.

2. La capa lógica donde residen los componentes centrales del sistema descritos anteriormente (NER, clustering, reglas y patrones), así como el propio motor de la aplicación.

3. Una subcapa de persistencia, que gestiona los datos y el acceso a las bases de datos del propio sistema.

4. Una capa de conectores, para interoperar con servicios y fuentes externas.

5. Las capas de seguridad y auditoría transversales.

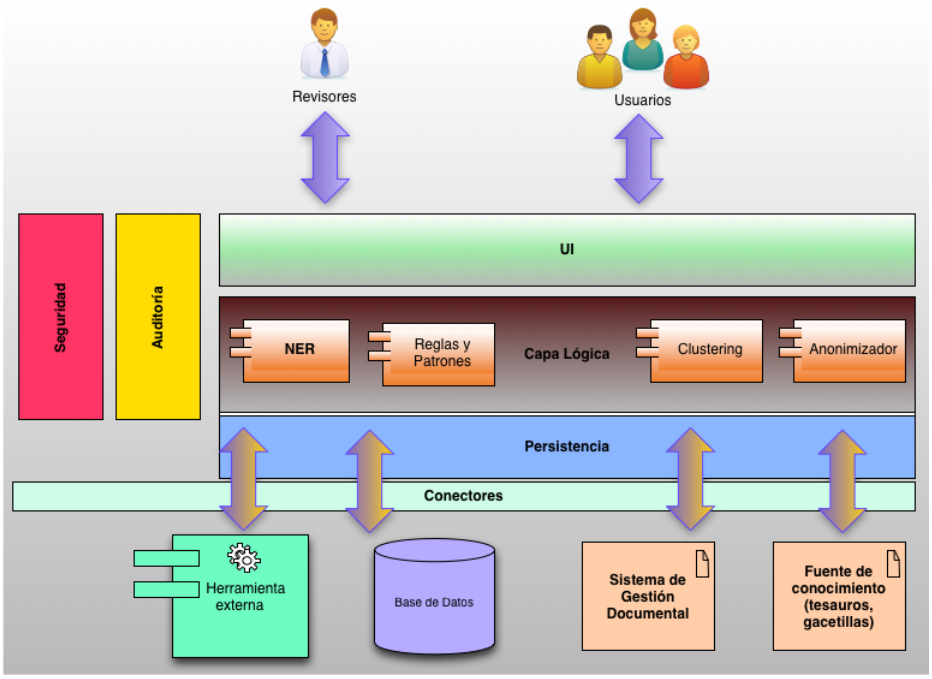

Figura 2: Modelo en capas del sistema

5.1.2. Modelado como proceso de negocios utilizando BPMN

A partir de la especificación del sistema descrita anteriormente, se visualizó que esta arquitectura particular permitía modelar el sistema como un proceso de negocios, mediante notación BPMN2[10].

El proceso es uno de los conductores de la arquitectura, y se presentan algunos requerimientos donde los motores de procesos y los BPMNS (BPM Suite) son especialmente útiles. Por BPMS nos referimos a un sistema integral para la implementación de procesos de negocio sobre el estándar BPMN2, que permite además de modelar, implementar y ejecutar los procesos, el modelado de la GUI, aportando infraestructura de seguridad y auditoría embebidas entre otros elementos.

En primer lugar vemos que el proceso de anonimización descrito, puede ser perfectamente modelado como un proceso de negocios, con tareas y transiciones bien definidas. De esta manera, se aporta entonces la flexibilidad y la potencia del modelado de procesos de negocios que permite adaptar con facilidad el proceso frente a distintos escenarios, como se puede apreciar en la Figura 3.

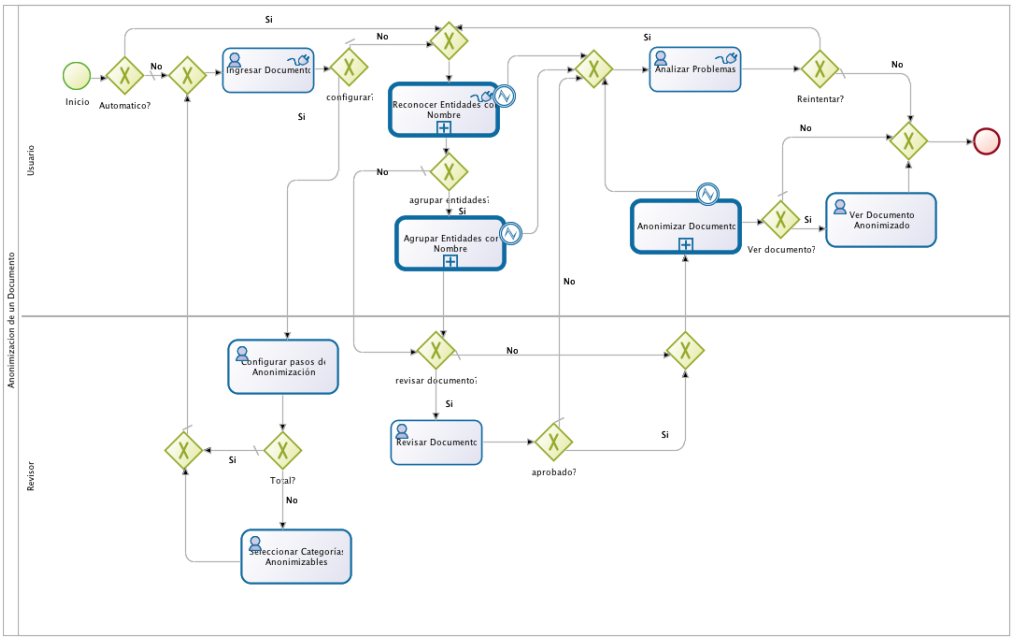

Figura 3: Proceso modelado mediante BPMN

La propuesta de modelado como proceso de negocios[7] constituye una característica específica de la arquitectura propuesta, que la diferencia con las arquitecturas estudiadas.

- El modelado del proceso mediante BPMN2[10], permite que se pueda personalizar el mismo, por ejemplo habilitando o deshabilitando ciertas tareas de acuerdo a la necesidad, sin que ello derive finalmente en una implementación o programación. Por otra parte definir el proceso mediante este estándar de la industria, está alineado con el principio de interoperabilidad definido como lineamiento general para ésta arquitectura.

- Tratándose de una arquitectura genérica, y teniendo en cuenta el requerimiento de adaptabilidad definido (RNF1), es necesario proveer mecanismos para que el proceso aquí descrito pueda adaptarse a dominios específicos y sistemas concretos. Una vez más la flexibilidad de BPMN2 permite atender esta necesidad.

- Se presentan requerimientos de auditoría y seguridad transversales, con necesidad de asignación de roles para distintantes tareas. Estos elementos son parte integral de las BPMS, y por ende se obtiene una notable ganancia en productividad a la hora de desarrollar un sistema de anonimización, ya que se podrá utilizar la infraestructura de seguridad y auditoría provista por el BPMS.

Pero quizás la mayor ganancia de utilizar BPMN2 es que una vez definido en este lenguaje, el proceso se puede pasar rápidamente del diseño a la ejecución sobre alguno de los motores de procesos de las BPMS existentes compatibles con este estándar. De esta manera toda la arquitectura presentada en la Figura 2 queda inmersa en el BPMS, a excepción de los componentes y sistemas externos que se pretendan integrar, tales como las que figuran en la zona inferior de dicha figura (herramientas externas, base de datos, sistemas de gestión documental y fuentes de conocimiento).

Finalmente, cabe destacar que de esta manera el proceso se vuelve extensible con facilidad, ya que se pueden incorporar nuevos elementos (tareas, subprocesos) nuevamente sin impacto en el diseño.

A continuación se describen las entradas, salidas y responsabilidades de los subprocesos y tareas del proceso.

Tareas

Nombre	**Ingresar Documento**
Responsabilidades	Esta tarea inicializa el documento a ser anonimizado.
Entrada	Un documento a anonimizar
Salida	Documento a ser anonimizado inicializado en el sistema.

Nombre	**Configurar pasos de la anonimización**
Responsabilidades	Esta tarea gestionará toda la configuración del sistema. Establece los parámetros de configuración que son utilizados por los distintos componentes
Entrada	Documentos seleccionados para anonimizar
Salida	Variables y parámetros de configuración establecidos.

Nombre	**Seleccionar Categorías Anonimizables**
Responsabilidades	Configuración complementaria. De tratarse de un proceso de anonimización parcial, permite seleccionar las categorías de Entidades con Nombre a ser anonimizadas.
Entrada	Variable “anonimización parcial” establecida.
Salida	Lista de categorías de Entidades con Nombre a anonimizar.

Nombre	**Reconocer Entidades Con Nombre (Subproceso)**
Responsabilidades	Este componente es el subproceso encargado de identificar las Entidades con Nombre en el texto.
Entrada	Un documento a anonimizar.
Salida	Lista de Entidades con Nombre identificadas en el documento.

Nombre	**Agrupar Entidades con Nombre (Subproceso)**
Responsabilidades	Este componente agrupa las entidades con nombre equivalentes en clusters.
Entrada	Documento con sus Entidades con Nombre identificadas.
Salida	Las entidades con nombre equivalentes agrupadas en clusters.

Nombre	**Revisar Documento**
Responsabilidades	Aprobar o rechazar la identificación de Entidades con Nombre realizada.
Entrada	Documento con sus Entidades con Nombre identificadas y agrupadas.
Salida	Documento aprobado o rechazado.

Nombre	**Anonimizar Documento (Subproceso)**
Responsabilidades	Este componente tiene la responsabilidad de remplazar la información sensible con datos genéricos o versiones cifradas de los datos, según la configuración del sistema.
Entrada	Documento aprobado, con sus Entidades con Nombre identificadas y agrupadas.
Salida	Documento anonimizado.

Nombre	**Ver Documento Anonimizado**
Responsabilidades	Mostrar el documento anonimizado al usuario.
Entrada	Documento anonimizado.
Salida	Fin del proceso.

Parámetros de Configuración

El proceso principal estará condicionado por una serie de variables y parámetros para representar las distintas configuraciones que son admitidas por el sistema. En el Cuadro 1 se detallan dichos parámetros.

Tabla 1: Parámetros

Nombre	Tipo	Parámetro	Método de Introducción
Texto	model.Text	El objeto documento a anonimizar. El tipo de datos model.Text será detallado en la Vista de Información más adelante en éste documento.	Interfaz con otro sistema, base de datos o introducción directa por parte del usuario.
Aprobado	Booleano	Variable para representar la aprobación o rechazo del documento	Establecida por el usuario con un checkbox en la tarea "Revisar Documento"
Automático	Booleano	Define si el proceso de anonimización será automático o será supervisado.	Establecida por el usuario con un checkbox en la tarea "Configurar pasos de la anonimización"
Agrupar	Booleano	Define si se deberán agrupar (clustering) las entidades con nombre.	Establecida por el usuario con un checkbox en la tarea "Configurar pasos de la anonimización"
Reversible	Booleano	Define si la anonimización será reversible o irreversible.	Establecida por el usuario con un checkbox en la tarea "Configurar pasos de la anonimización"
herramientaNER	Texto	Nombre de la herramienta NER que se utilizará.	Seleccionable por el usuario en un combo box en la tarea "Configurar pasos de la anonimización"
herramientaClustering	Texto	Nombre de la herramienta de clustering que se utilizará.	Seleccionable por el usuario en un combo box en la tarea "Configurar pasos de la anonimización"
heuristicas	Booleano	Define si se procesará el documento mediante reglas heurísticas, reglas y patrones.	Establecida por el usuario con un checkbox en la tarea "Configurar pasos de la anonimización"
Total	Booleano	Define si la anonimización será parcial (false) o total (true).	Establecida por el usuario con un checkbox en la tarea "Configurar pasos de la anonimización"
Otra_herramienta	Booleano	Define si se deberán invocar herramientas externas adicionales.	Establecida por el usuario con un checkbox en la tarea "Configurar pasos de la anonimización"

Dentro del proceso de anonimización, se identificaron los subprocesos que se describen a continuación.

Subproceso Reconocer Entidades con Nombre

Este subproceso se ilustra en la Figura 4, y tiene como cometido identificar las entidades con nombre. Como tareas del proceso destacan la invocación de herramientas NER, herramienta externas (por ejemplo tesauros, gacetillas, etc), y la invocación de motores de reglas y patrones.

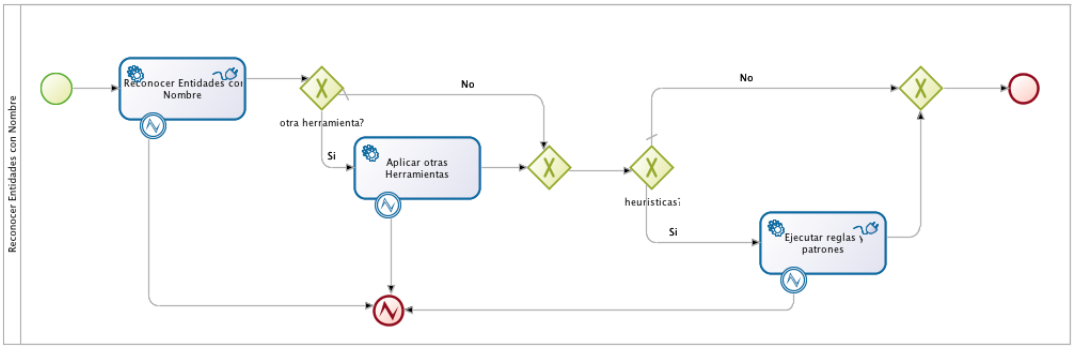

Figura 4: Subproceso Reconocer Entidades con Nombre

Dentro de este subproceso, se encuentra una primera tarea que es la que efectivamente invoca alguna herramienta con capacidades NER concretas. Esta tarea se puede encontrar en las arquitecturas estudiadas. Por ejemplo, es identificada como NE Recognition en MOSTAS[2] y como Tagger en ANONIMYTEXT[11].

Las tareas adicionales del subproceso posibilitan la ejecución de herramientas complementarias para mejorar la identificación de entidades con nombre. En las propuestas estudiadas también se contemplaba la integración de herramientas y servicios adicionales, tales como gacetillas, diccionarios de acrónimos, correctores ortográficos, etc, por tanto esta propuesta deja abierta esta posibilidad.

Tareas

Nombre	**Reconocer Entidades con Nombre**
Responsabilidades	Invoca a la herramienta NER
Entrada	Documento
Salida	Documento con Entidades con Nombre identificadas.

Nombre	Aplicar otras herramientas
Responsabilidades	Este servicio tiene la responsabilidad de manejar la interoperabilidad del sistema con fuentes externas de conocimiento.
Entrada	Documento con Entidades con Nombre identificadas.
Salida	Documento con Entidades con Nombre identificadas (con posibles mejoras en la identificación)

Nombre	Ejecutar reglas y patrones
Responsabilidades	Se aplican reglas heurísticas y patrones para identificar Entidades con Nombre adicionales.
Entrada	Documento con Entidades con Nombre identificadas.
Salida	Documento con Entidades con Nombre identificadas (con posibles mejoras en la identificación)

Subproceso Agrupar Entidades con Nombre

Las entidades con nombre identificadas son clasificadas y agrupadas en clusters, mediante algún algoritmo de agrupamiento. El objetivo es que referencias a una misma entidad sean agrupadas. Por ejemplo, "Naciones Unidas" es lo mismo que "O.N.U." El subproceso se muestra en la Figura 5.

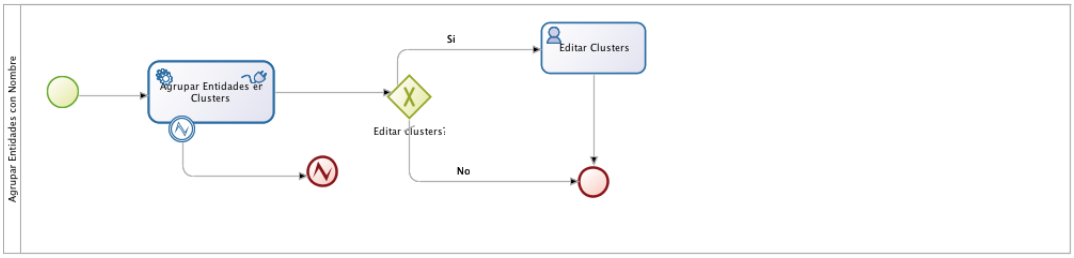

Figura 5: Subproceso Agrupar Entidades con Nombre

Tareas

Nombre	Agrupar Entidades en Clusters
Responsabilidades	Invoca a herramientas de clustering, agrupando las Entidades con Nombre equivalentes.
Entrada	Documento con Entidades con Nombre identificadas.
Salida	Documento con Entidades con Nombre identificadas y agrupadas.

Nombre	Editar Clusters
Responsabilidades	Esta tarea permite al usuario corregir y editar los grupos (clusters) identificados.
Entrada	Documento con Entidades con Nombre identificadas y agrupadas.
Salida	Documento con Entidades con Nombre identificadas y agrupadas (con posibles mejoras en la agrupación)

Subproceso Anonimizar Documento

Estando identificadas las entidades con nombre, este subproceso se encarga de la anonimización en si misma. Se identificaron distintos tipos de anonimización, parcial o total según si se anonimizan todos los tipos de entidades con nombre, reversible o irreversible si un documento anonimizado puede revertirse al original o no respectivamente. El subproceso se ilustra en la Figura 6.

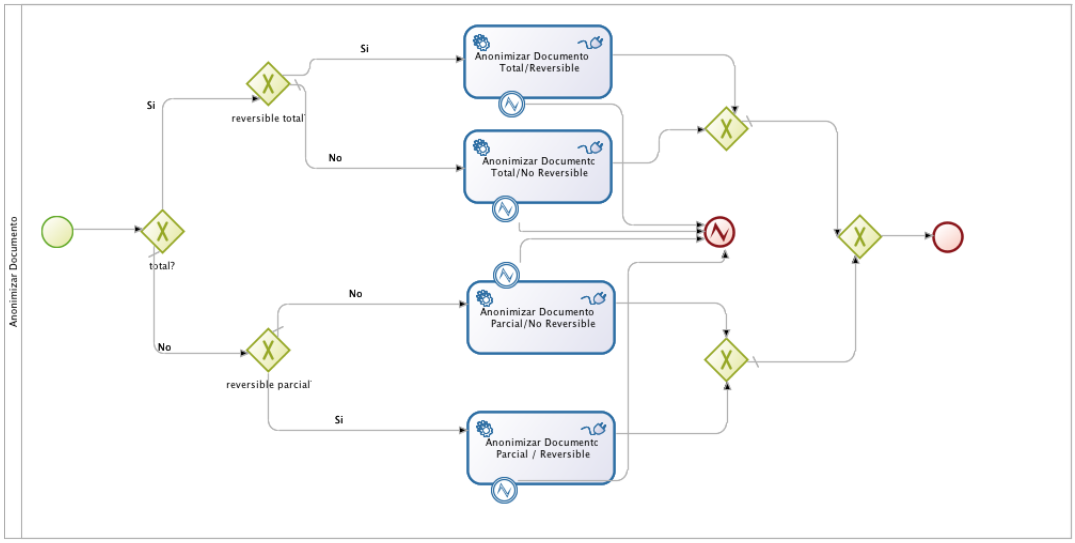

Figura 6: Subproceso Anonimizar Documento

Tareas

Nombre	Anonimizar Documento Total/Reversible
Responsabilidades	Sustituye todas las Entidades con Nombre identificadas en el documento por una referencia cifrada reversible.
Entrada	Documento con sus Entidades con Nombre identificadas.
Salida	Documento anonimizado

Nombre	Anonimizar Documento Total/No Reversible
Responsabilidades	Sustituye todas las Entidades con Nombre identificadas en el documento por un texto genérico.
Entrada	Documento con sus Entidades con Nombre identificadas y agrupadas.
Salida	Documento anonimizado

Nombre	Anonimizar Documento Parcial/No Reversible
Responsabilidades	Sustituye las Entidades con Nombre que pertenezcan a las categorías configuradas, por un texto genérico.
Entrada	Documento con sus Entidades con Nombre identificadas y agrupadas.
Salida	Documento anonimizado.

Nombre	Anonimizar Documento Parcial / Reversible
Responsabilidades	Sustituye las Entidades con Nombre que pertenezcan a las categorías configuradas, por una referencia cifrada reversible.
Entrada	Documento con sus Entidades con Nombre identificadas y agrupadas.
Salida	Documento anonimizado.

5.2. Vista de Información

Esta sección presenta la Vista de Información, mediante la cual se describe cómo se representa el modelo de datos del sistema de anonimización, y el formato en el cual fluye la información entre los distintos componentes.

5.2.1. Estructura de Datos

Para el flujo del texto a ser anonimizado dentro del sistema, se propone una estructura de datos sencilla que se aprecia en la Figura 7, consistente en una entidad que denominaremos Text que contendrá la cadena (String) conteniendo el documento en su estado actual (texto original o anonimizado), y una estructura conteniendo el conjunto de Entidades con Nombre identificadas para el documento.
La entidad con nombre se modela con una clase específica denominada NamedEntity, también especifica en la Figura 7, que tiene tres atributos:

1. term: Cadena que contiene el término o nombre. Ejemplo: “Juan Pérez”.

2. neClass: Cadena que contiene el tipo de entidad con nombre clasificada: Ejemplo: “ORGANIZATION”, “GEO_LOCATION”, “PERSON”

3. aliases: Lista de términos equivalentes a la entidad con nombre definida en term. Aquí se guardan las entidades equivalentes al agrupar las entidades (clustering)

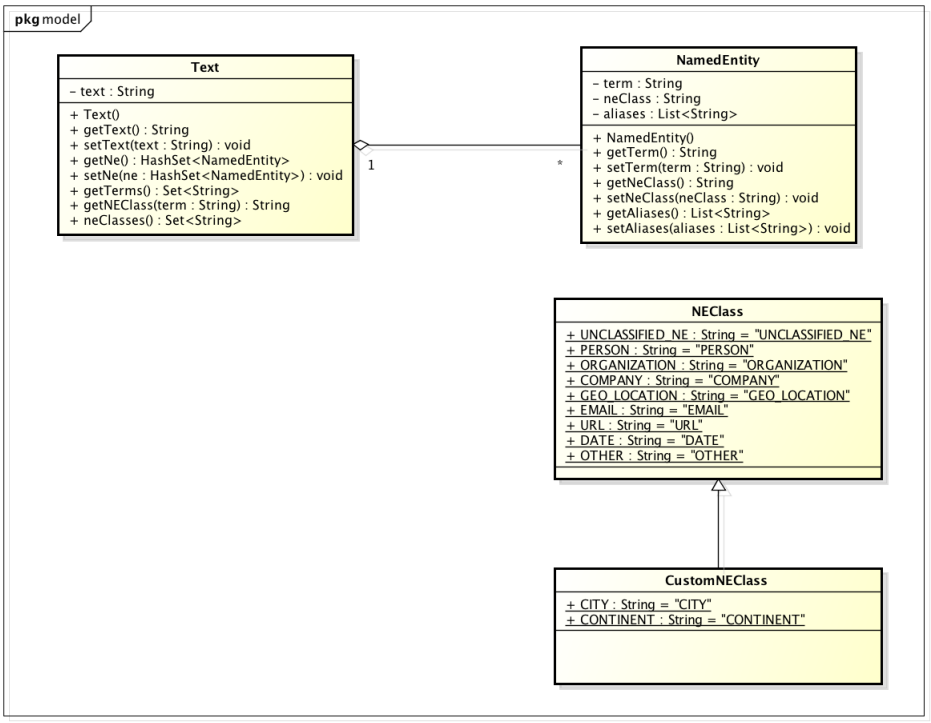

Figura 7: Modelo de Datos

5.2.2. Flujo de Datos

La información fluye en el sistema en el formato de un documento de texto presentada en la sección anterior. Si bien al haberse modelado el proceso utilizando notación BPMN el flujo resulta autoexplicativo, en la Figura 8 se ilustra de una forma alternativa el flujo de información. Allí se puede apreciar como el documento pasa por los distintos componentes del sistema de manera secuencial.

1. En primer lugar se procesa el documento mediante el componente NER, el cual realiza la identificación de entidades con nombre primaria.

2. Seguidamente el documento es procesado mediante las reglas y patrones heurísticos que complementan al proceso NER anterior.

3. Se agrupan las entidades con nombre equivalente mediante técnicas de clustering. Se presenta el agrupamiento, y en caso de existir errores se retroalimenta al sistema y se reinicia el proceso.

4. El documento pasa al módulo Revisor, que permite a un experto validar el resultado de la identificación de entidades con nombre. En caso de no aprobación se retroalimenta al sistema y se reinicia desde el módulo NER.

5. Finalmente el documento es procesado por el módulo de Anonimización, el cual sustituye o cifra las entidades con nombre en el documento.

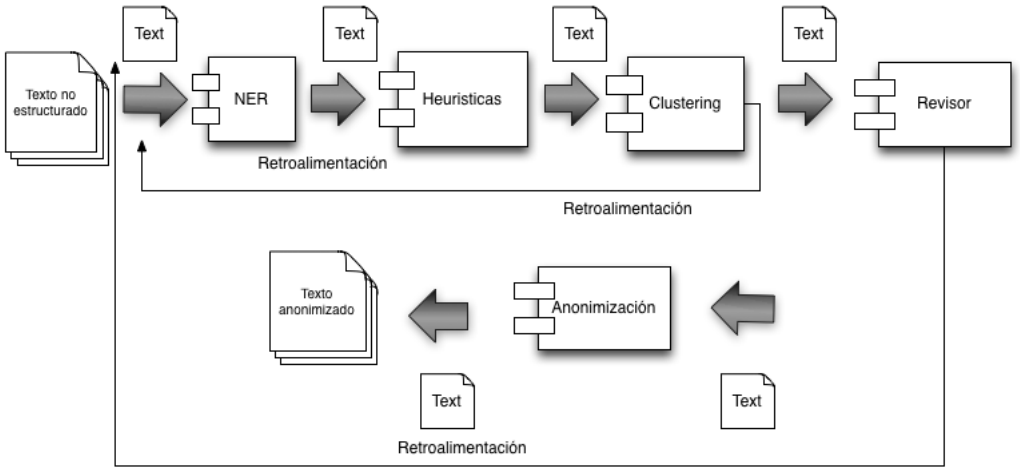

Figura 8: Flujo de información

5.3. Vista de Desarrollo

La presente Vista de Desarrollo, describe organización modular del sistema (estructura de paquetes), y los estándares de diseño que se utilizan en el sistema de anonimización.

5.3.1. Estructura de Paquetes

En la figura Estructura de Paquetes se describe una estructura de paquetes genérica para organizar los servicios que deberán ser implementados e invocados desde el flujo de trabajo del sistema.

Como se aprecia en la figura, se proponen tres paquetes para organizar los diferentes módulos del sistema:

1. **Paquete anonimization.externaltoolwrapper:** Este paquete contiene los módulos NER, heurísticas y clustering, cada uno en un subpaquete específico. El común denominador es la interacción con herramientas externas que realizan cada uno de estos procesamientos de texto, mediante la utilización de adaptadores.

2. **Paquete anonimization.model:** En este paquete se concentra el modelo de datos que se maneja en el sistema, las clases Text, NamedEntity y NEClass, que fueran descritas en la Vista de Información precedente en este documento.

3. **Paquete anonimization.anonymizer:** Aquí se define el módulo Anonimizador del sistema, el cual será el encargado de cifrar o sustituir las entidades con nombre en el documento. Se define una clase Anonymizer, especializada por dos clases, una para representar un anonimizador reversible (ReversibleAnonimizer), que sustituye las entidades con nombre por el propio dato pero luego de procesarlo con un cifrado reversible, y un anonimizador no reversible (NonReversibleAnonymizer), el cual simplemente sustituye las entidades con nombre por información genérica.

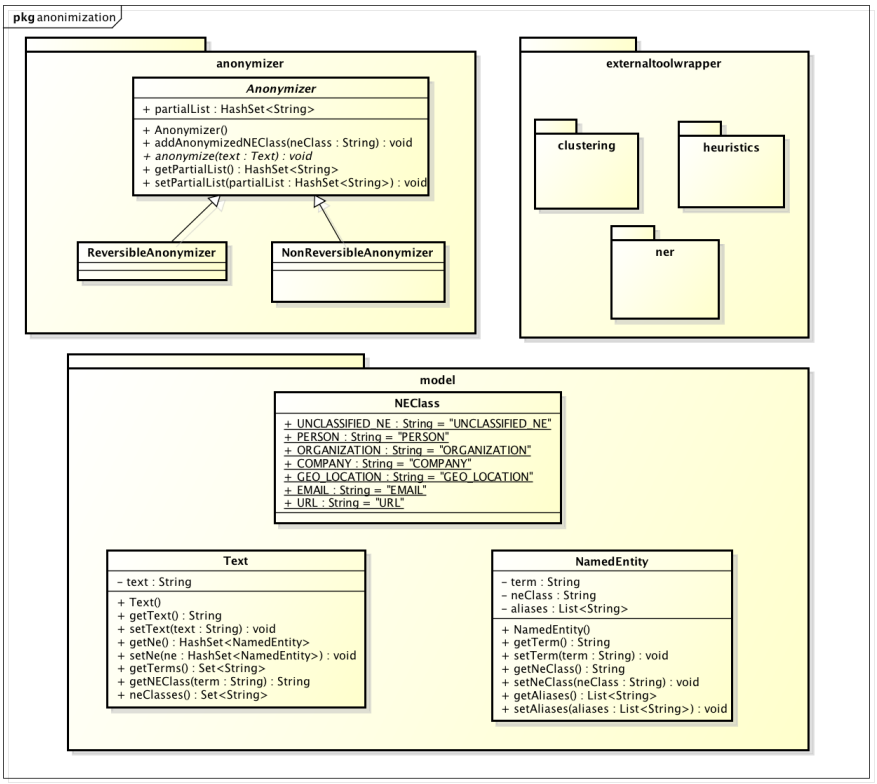

Figura 9: Estructura de Paquetes

5.3.2. Estándares de diseño

El sistema presenta algunos requerimientos clásicos, que son resueltos por aplicación de patrones de diseño. Tal estrategia es aplicada en los componen-

tes del paquete “externaltoolwrapper.ner”, donde se utiliza el patrón de diseño Adapter, como se ilustra en la Figura 10, para modelar la interacción del sistema con herramientas externas, cuyas interfaces se desea abstraer. Para ello se define una superclase con métodos abstractos que deberán ser implementados por las especializaciones.

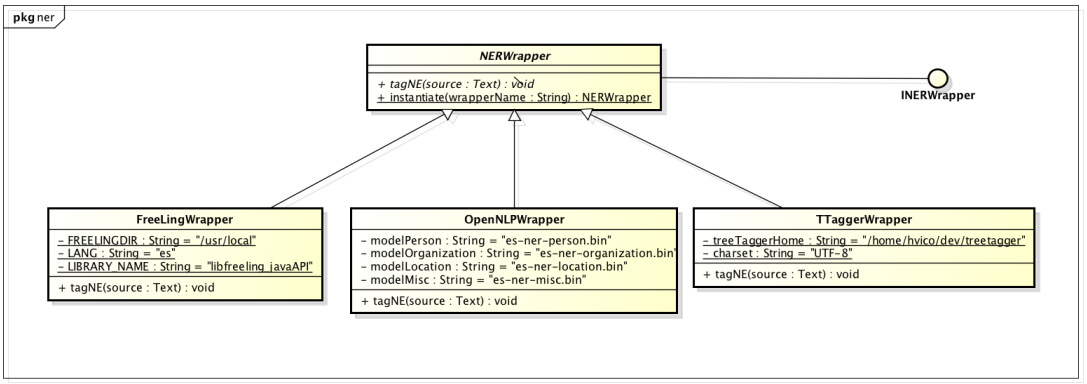

Figura 10: Patrón Adapter

6. Respuesta de la Arquitectura a los Requerimientos

En esta sección se presentará un resumen de los requerimientos, y cómo la arquitectura propuesta los satisface.

6.1. Requerimientos funcionales

Req.	**Descripción del Requerimiento**
RF1	El sistema debe poder procesar documentos no estructurados, en formato texto plano.
Respuesta de la Arquitectura	
La arquitectura propuesta contempla el procesamiento de documentos no estructurados, y los representa internamente mediante la estructura "Text" presentada en la Vista de Información.	

Req.	**Descripción del Requerimiento**
RF2	Se deben poder utilizar distintas herramientas de software para el procesamiento del texto indistintamente. La herramienta a utilizar en cada etapa debe ser un elemento más de configuración.
Respuesta de la Arquitectura	
Se provee una capa de conectores, con adaptadores para integrar distintas herramientas de procesamiento de texto. El proceso es configurable y permite seleccionar la herramienta a utilizar.	

Req.	**Descripción del Requerimiento**
RF3	El sistema debe permitir a un experto del dominio validar la salida anonimizada, y permitir aprobar el documento o reprobarlo brindando feedback que retroalimente al sistema.
Respuesta de la Arquitectura	
El proceso definido contempla una tarea específica para revisión, y la consecuente aprobación o rechazo de los documentos. Esta tarea se define de forma específica para el rol Revisor.	

Req.	Descripción del Requerimiento
RF4	Se debe poder configurar el alcance de la anonimización, la cual puede ser total (todos los atributos que identifican a personas u organizaciones), o parcial (un subconjunto de los atributos).
Respuesta de la Arquitectura	
El proceso principal contempla la selección de las categorías anonimizables en la tarea "Seleccionar Categorías Anonimizables". El subproceso Anonimizar Documento contempla la anonimización parcial o total en base de acuerdo al parámetro "Total" definido para el proceso principal. La superclase Anonymizer del paquete anonymizer provee una estructura para definir el subconjunto de categorías a anonimizar en caso de una anonimización parcial.	

Req.	Descripción del Requerimiento
RF5	El sistema debe permitir por configuración, realizar una anonimización irreversible (se elimina por completo la información sensible), o reversible (se sustituye la información sensible por una referencia cifrada a la misma).
Respuesta de la Arquitectura	
El proceso principal contempla un parámetro de configuración para definir si la anonimización es reversible o irreversible. Asimismo el subproceso anonimización de un documento define tareas específicas de acuerdo a dicho parámetro de configuración, aplicando anonimización reversible o irreversible según corresponda. En el paquete anonymizer existen especializaciones del componente anonimizador para contemplar ambos casos.	

Req.	Descripción del Requerimiento
RF6	El sistema debe permitir la introducción de reglas o patrones (heurísticas) definidas por un usuario experto, para contemplar características específicas de los documentos de un dominio en particular.
Respuesta de la Arquitectura	
Se provee un módulo para procesar reglas y patrones, dentro del paquete externaltoolwrapper.heuristics. El subproceso "Reconocer entidades con nombre" contempla una tarea específica para invocar a dicho componente.	

Req.	Descripción del Requerimiento
RF7	El sistema debe contemplar la posibilidad de interactuar con fuentes externas, mediante adaptadores, que provean información de términos específicos de un dominio (diccionarios, tesauros, gacetillas, acrónimos, etc.)
Respuesta de la Arquitectura	
El subproceso "Reconocer entidades con nombre" contempla una tarea específica para invocar herramientas externas. El diseño mediante aplicación del patrón "Adaptador" (wrapper) facilita la integración de herramientas específicas.	

Req.	Descripción del Requerimiento
RF7	El sistema debe contemplar la posibilidad de interactuar con fuentes externas, mediante adaptadores, que provean información de términos específicos de un dominio (diccionarios, tesauros, gacetillas, acrónimos, etc.)
Respuesta de la Arquitectura	
El subproceso "Reconocer entidades con nombre" contempla una tarea específica para invocar herramientas externas. El diseño mediante aplicación del patrón "Adaptador" (wrapper) facilita la integración de herramientas específicas.	

6.2. Requerimientos no funcionales

Req.	Descripción del Requerimiento
RNF1	Adaptabilidad
Respuesta de la Arquitectura	
Se diseña el proceso utilizando BMPNv2. Esto permite que dicho proceso pueda adaptarse con facilidad en base a requerimientos y necesidades específicas. La aplicación del patrón de diseño "Adaptador" (wrapper) facilita la integración de herramientas específicas.	

Req.	Descripción del Requerimiento
RNF2	Configurabilidad
Respuesta de la Arquitectura	
El proceso contempla tareas específicas tendientes a definir la configuración del sistema. Se define un conjunto de parámetros de configuración, los cuales pueden ser establecidos por el usuario.	

Req.	Descripción del Requerimiento
RNF3	Interoperabilidad
Respuesta de la Arquitectura	
El subproceso "Reconocer entidades con nombre" contempla una tarea específica para invocar herramientas externas. El diseño mediante aplicación del patrón "Adaptador" (wrapper) facilita la integración de herramientas específicas.	

Req.	Descripción del Requerimiento
RNF4	Documentación: Las interfaces del sistema deberán estar debidamente documentadas, de forma que un usuario avanzado pueda agregar nuevas fuentes de conocimiento o herramientas de procesamiento de lenguaje natural.
Respuesta de la Arquitectura	
El presente documento propone la aplicación del patrón adaptador para la integración de las herramientas de procesamiento de lenguaje natural. Se describe y documento un modelo de ejemplo en la Vista de Desarrollo.	

Req.	Descripción del Requerimiento
RNF5	Extensibilidad: El añadir nuevas herramientas de procesamiento de lenguaje (RF2 y RNF1) deberá ser un proceso sencillo, que no implique modificaciones mayores al sistema.
Respuesta de la Arquitectura	
El subproceso "Reconocer entidades con nombre" contempla una tarea específica para invocar herramientas externas. El diseño mediante aplicación del patrón "Adaptador" (wrapper) facilita la integración de herramientas específicas.	

Req.	Descripción del Requerimiento
RNF6	Auditoría: El sistema deberá registrar las interacciones de usuarios con los documentos.
Respuesta de la Arquitectura	
Estando modelado el sistema como un proceso de negocios BPMNv2, se puede ejecutar el proceso sobre un motor que soporte dicha especificación, obteniendo de forma automática la trazabilidad de las tareas ejecutadas por cada usuario.	

Req.	Descripción del Requerimiento
RNF7	Seguridad: Los documentos no anonimizados no deben ser accedidos por usuarios no autorizados.
Respuesta de la Arquitectura	
Tal como ocurre con la auditoría, utilizando motores BPMNv2 se obtienen "de facto" la capa de autenticación y gestión de la seguridad propia de dichas soluciones.	

7. Apéndices

7.1. Apéndice: Decisiones y Alternativas

Una propuesta de arquitectura alternativa que se consideró, consiste en la aplicación del patrón de arquitectura "Blackboard" (pizarra). En este se pensaron los subprocesos actuando como "agentes" independientes, tomando y devolviendo el texto desde y hacia una estructura de datos común (pizarra). Se definiría además una estrategia, que sería la encargada de orquestar la interacción de los distintos agentes con la pizarra.
Esta propuesta se descartó porque de la forma en que trabajan las distintas herramientas de software vistas, para actuar como agentes independientes la estrategia terminaría prácticamente secuenciando de todas formas el trabajo de los agentes sobre un texto en particular. Varios módulos de los identificados como necesarios requieren de la finalización de subprocesos previos para activarse. Algunos de los subprocesos identificados en la anonimización, pueden resultar casos interesantes para la aplicación de patrones como blackboard, donde la solución no es determinista y la verificación de la finalización del proceso resulta compleja. Sin embargo la idea de ésta arquitectura de referencia es "encapsular" estos subprocesos complejos, los cuales como se vio en la etapa de investigación del estado del arte son problemas ya estudiados y/o "resueltos" tanto desde el punto de vista teórico/académico como en su implementación mediante herramientas de software.
Por tal motivo al estar trabajando a un nivel de abstracción superior al de los algoritmos de procesamiento de lenguaje natural, se ve la alternativa de un proceso basado de "pipes and filters" una aproximación más adecuada al problema, que además sigue la línea de todas las propuestas de arquitectura estudiadas.

Referencias

[1] Gutiérrez A. Extracción de entidades con nombre ó named entity recognition (ner). URL http://www.oocities.org/es/extracciondeentidadesdenombre/.

[2] Iglesis A., Castro R., Pérez L., Castaño P., Martínez J., Gómez-Pérez S., and Melero R. 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.

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

[4] República Oriental del Uruguay Poder Legislativo. 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.

[5] Buschmann F., Meunier R., Rohnert H., Sommerlad P., and Stal M. *Pattern-Oriented Software Architecture, Volume 1: A System of Patterns*. Wiley, Chichester, UK, 1996. ISBN 978-0-471-95869-7.

[6] Natividad Prieto Ferran Pla, Antonio Molina. Evaluación de un etiquetador morfosintáctico basado en bigramas especializados para el castellano. *Procesamiento de Lenguaje Natural*, 27(0), 2001. ISSN 1989-7553. URL http://sinai.ujaen.es/sepln/ojs/ojs/index.php/pln/article/view/3362.

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

[8] 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.

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

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

[11] Perez-Lainez R, De Pablo-Sanchez C., and Iglesias A. *ANONIMYTEXT : Anonymization of unstructured documents*. 2008.

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

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