Osteba

OSASUN TEKNOLOGIEN EBALUAZIOA

EVALUACIÓN DE TECNOLOGÍAS SANITARIAS

EUSKO JAURLARITZA
GOBIERNO VASCO

OSASUN ETA KONTSUMO SAILA
DEPARTAMENTO DE SANIDAD Y CONSUMO

INFORME DE EVALUACIÓN

D-11-05

DESARROLLO DEL CONJUNTO DE DATOS BÁSICOS DE LA ASISTENCIA AMBULATORIA ESPECIALIZADA (CDB-AAE) DE OSAKIDETZA. EXPERIENCIA PILOTO EN TRES HOSPITALES

Proyecto de Investigación Comisionada

Marzo 2011

Osteba

INFORME DE EVALUACIÓN

D-11-05

DESARROLLO DEL CONJUNTO DE DATOS BÁSICOS DE LA ASISTENCIA AMBULATORIA ESPECIALIZADA (CDB-AAE) DE OSAKIDETZA. EXPERIENCIA PILOTO EN TRES HOSPITALES

Proyecto de Investigación Comisionada

Marzo 2011

Javier Yetano Laguna
José Manuel Ladrón de Guevara Portugal
Gonzalo López Arbeloa
Julián Salvador Blanco
Santiago Rodríguez Tejedor
Mikel Ogueta Lana
Jon Guajardo Remacha

EUSKO JAURLARITZA GOBIERNO VASCO

OSASUN ETA KONTSUMO SAILA
DEPARTAMENTO DE SANIDAD Y CONSUMO

Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia
Servicio Central de Publicaciones del Gobierno Vasco

Vitoria-Gasteiz, 2011

Un registro bibliográfico de esta obra puede consultarse en el catálogo de la Biblioteca General del Gobierno Vasco: <http://www.euskadi.net/ejgvbiblioteka>

Financiación: Beca de Investigación Comisionada 2007. Departamento de Sanidad. Gobierno Vasco. N.º Expediente 2007/01.

Este documento debe ser citado como:

Yetano Laguna J, Ladrón de Guevara Portugal JM, López Arbeloa G, Salvador Blanco J, Rodríguez Tejedor S, Ogueta Lana M, Guajardo Remacha J. *Desarrollo del Conjunto de Datos Básicos de la Asistencia Ambulatoria Especializada (CDB-AAE) de Osakidetza. Experiencia piloto en tres hospitales*. Investigación Comisionada. Vitoria-Gasteiz. Departamento de Sanidad, Gobierno Vasco, 2011. Informe n.º: Osteba D-11-05.

El contenido de este documento refleja exclusivamente la opinión de las personas investigadoras, y no son necesariamente compartidas en su totalidad por quienes han realizado la revisión externa o por el Departamento de Sanidad y Consumo del Gobierno Vasco.

Edición:	1.ª julio 2011
Tirada:	80 ejemplares
©	Administración de la Comunidad Autónoma del País Vasco Departamento de Sanidad y Consumo
Internet:	www.euskadi.net/sanidad/osteba
Edita:	Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia Servicio Central de Publicaciones del Gobierno Vasco Donostia-San Sebastián, 1 - 01010 Vitoria-Gasteiz
Fotocomposición:	Ipar, S. Coop. Zurbaran, 2-4 – 48007 Bilbao
Impresión y encuadernación:	Grafo, S.A. Avda. Cervantes, 51 – 48970 Basauri (Bizkaia)
ISBN:	978-84-457-3160-4
D.L.:	BI 2.077-2011

Investigador principal

Javier Yetano Laguna. Servicio de Documentación Clínica. Hospital de Galdakao-Usansolo. Galdakao (Bizkaia).

Miembros del equipo de investigación

José Manuel Ladrón de Guevara Portugal. Subdirección de asistencia especializada. Organización Central de Osakidetza. Vitoria-Gasteiz (Álava).

Gonzalo López Arbeloa. Subdirección de Calidad. Organización Central de Osakidetza. Vitoria-Gasteiz (Álava).

Julián Salvador Blanco. Dirección Médica. Hospital Donostia. Donostia-San Sebastián (Gipuzkoa).

Santiago Rodríguez Tejedor. Servicio de Documentación Clínica. Hospital de Cruces. Barakaldo (Bizkaia).

Mikel Ogueta Lana. Subdirección de asistencia especializada. Organización Central de Osakidetza. Vitória-Gasteiz (Álava).

Jon Guajardo Remacha. Dirección Médica. Hospital de Galdakao-Usánsolo. Galdakao (Bizkaia).

Revisores externos

Dr. Pablo Arbeloa López. Innovasalud. Santander (Cantabria).

Dr. Orencio López Domínguez. Innovasalud. Santander (Cantabria).

Coordinación del Proyecto en Osteba

Asun Gutierrrez Iglesias. Servicio de Evaluación de Tecnologías Sanitarias, Osteba. Departamento de Sanidad y Consumo. Gobierno Vasco. Vitoria-Gasteiz (Álava).

ÍNDICE

RESÚMENES ESTRUCTURADOS

RESUMEN ESTRUCTURADO

Título: DESARROLLO DEL CONJUNTO DE DATOS BÁSICOS DE LA ASISTENCIA AMBULATORIA ESPECIALIZADA (CDB-AAE) DE OSAKIDETZA. EXPERIENCIA PILOTO EN TRES HOSPITALES.

Autores: Yetano J, Ladrón de Guevara JM, López J, Salvador J, Rodríguez S, Ogueta M, Guajardo J.

Tecnología: Organización

Palabras clave MeSH: *CMBD, Inteligencia artificial, Codificación artificial, CIE-9-MC*

Fecha: Marzo 2011

Páginas: 58

Referencias: 12

Lenguaje: Castellano, resúmenes en inglés y euskera

ISBN: 978-84-457-3160-4

INTRODUCCIÓN

En los últimos años la asistencia ambulatoria tiene cada vez más peso en nuestros hospitales. En 2009 se produjeron 3.880.141 consultas y 912.221 urgencias hospitalarias en los hospitales de agudos de Osakidetza y cada día crece más la asistencia en el hospital de día. A pesar de ello, no se dispone de un conjunto de datos normalizado en el área ambulatoria similar al CMBD de hospitalización que ha demostrado ser muy útil. No se tiene información de qué enfermedades tienen los enfermos atendidos ambulatoriamente ni qué procedimientos se les realizaron. Por ello, no se tiene conocimiento de la casuística de cada centro o servicio médico o médico concreto ni de los aspectos relacionados con los resultados de esa asistencia para poder evaluar su eficiencia y su calidad. La mayor dificultad está en la recogida de los datos (y especialmente la de codificar los diagnósticos y los procedimientos) de cientos de miles de asistencias ambulatorias.

OBJETIVOS

Definir un Conjunto de Datos Básicos de la Asistencia Ambulatoria Especializada (CDB-AAE), desarrollar una aplicación que permita la recogida del CDB-AAE y la codificación automática de los diagnósticos y procedimientos y su implementación en tres hospitales.

MATERIAL Y MÉTODOS

Se ha revisado la literatura científica y la normativa de las diferentes comunidades autónomas, se ha definido un CDB-AAE con las variables a recoger, se ha evaluado la calidad de la codificación descentralizada de urgencias en 4 hospitales de Osakidetza y se ha desarrollado el programa KodifiKa (de codificación automática con la CIE-9-MC de los diagnósticos introducidos por el profesional que presta la asistencia ambulatoria).

KodifiKa trata una expresión literal de un diagnóstico y es capaz de codificarla en función del conocimiento contenido en una base de datos de literales de diagnósticos ya codificados utilizando inteligencia artificial en la valoración semántica de la expresión diagnóstica.

Análisis económico: SI (NO) **Opinión de expertos::** SI (NO)

RESULTADOS

Se presentan la valoración de la calidad de la codificación descentralizada en los episodios de urgencia, la lista de variables del CDB-AAE consensuada y los datos del uso de KodifiKa (80,3% de codificación automática) así como su descripción técnica en el Anexo 1.

CONCLUSIONES

Tener un CBD-AAE en nuestro Sistema Nacional de Salud es algo deseable y alcanzable si el Departamento de Sanidad lo normaliza, se extiende la Historia Clínica Electrónica y se resuelve el problema de la codificación de los diagnósticos y procedimientos de cantidades ingentes de asistencias ambulatorias con sistemas semiautomáticos que aseguren la calidad. Defendemos la implementación de un programa de codificación mixto (descentralizado-centralizado) como KodifiKa.

LABURPEN EGITURATUA

Izenburua: OSAKIDETZAKO LAGUNTZA ANBULATORIO ESPEZIALIZATUAREN OINARRIZKO DATU BILDUMAREN GARAPENA (ODB-LAE). ESPERIENTZIA PILOTUA HIRU OSPITALETAN

Egileak: Yetano J, Ladrón de Guevara JM, López J, Salvador J, Rodríguez S, Ogueta M, Guajardo J.

Teknologia: Antolaketa

MeSH gako-hitzak: *XOGB Inteligentzia artifiziala, Kodifikazio artifiziala, GNS-9-AK*

Data: 2011ko martxoa

Orrialdeak: 58

Erreferentziak: 12

Hizkuntza: Gaztelania, laburpenak ingelesez eta euskaraz

ISBN: 978-84-457-3160-4

HITZAURREA

Azken urteotan laguntza anbulatorioa pisu gutxiago hartzen ari da gure ospitaleetan. 2009an Osakidetzako akutuentzako ospitaleetan 3.880.141 kontsulta eta 912.221 ospitaleko urgentzia egon ziren. Halere, eremu anbulatorioan ez dago ospitaleratzeko DGOBren antzekoa den datu normalizatuen multzorik (datu normalizatuen multzo hori frogatuta dago baliagarria dela). Ez dago informaziorik anbulatorioetan artatuak izan diren gaixoen gaixotasunei buruz, ez eta erabili ziren prozedurei buruz ere. Horregatik, ez dugu ezagutzen zentro, zerbitzu mediko edo mediko bakoitzaren kasuistika, ez eta egindako artatzeen emaitzak, gerora horien efizientzia eta kalitatea ebaluatu ahal izateko. Zailtasunik handiena da ehunka mila laguntza anbulatorioen datu bilketa egitea (eta batez ere diagnostikoak eta prozedurak kodifikatzea).

HELBURUAK

Laguntza Anbulatorio Espezializatuaren Oinarrizko Datu multzoa definitzea (ODB-LAE). Aplikazio bat garatzea ondokoa ahalbideratzeko: ODB-LAEren bilketa eta diagnostikoen eta prozeduren kodifikazioa eta horien inplementazioa hiru ospitaleetan.

MATERIALAK ETA METODOAK

Literatura zientifikoa eta hainbat erkidego autonomoren araudia aztertu da. ODB-LAE bat definitu da, jaso beharreko aldagaiekin. Osakidetzako 4 ospitaletako urgentzietako kodifikazio deszentralizatuaren kalitatea ebaluatu da eta KodifiKa programa garatu da (kodifikazio automatikokoa; anbulatorioko artatzea egiten duen profesionalak sartutako diagnostikoen GNS-9-AKarekin).

KodifiKak diagnostiko baten hitzez hitzezko adierazpenak tratatzen ditu eta gauza da kodifikazioa egiteko kodifikatuta dauden diagnostikoen hitzez hitzeko datu basean dagoen ezagutzaren arabera, horretarako diagnostikatutako adierazpenaren balorazio semantikoa egiteko inteligentzia artifiziala erabilita.

Ekonomia-analisia: BAI (EZ) **Adituen iritzia:** BAI (EZ)

EMAITZAK

Ondokoak aurkezten dira: urgentziako episodioetako kodifikazio deszentralizatuaren kalitatearen balorazioa; adostutako ODM-LAEren adierazleen zerrenda; eta KodifiKaren erabilpen datuak (%80,3 kodifikazio automatikokoa) eta haren deskripzioa 1 Eranskinean.

ONDORIAK

Estatuko Osasun Sisteman ODM-LAE bat izatea desiragarria eta lorgarria da baldin eta Osasun Sailak normalizatzen badu, Historia Kliniko Elektronikoa zabaltzen bada eta ondoko arazoa konpontze bada: diagnostikoen eta laguntza anbulatorioen kopuru handien prozedurak kodifikatzeko kalitatea ziurtatuko duten sistema semiautomatikoak jartzen badira.

Kodifikazio mistoa (desentralizatua-zentralizatua) egiten duen KodifiKa bezalako programa bat inplementatzearen aldekoak gara.

STRUCTURED SUMMARY

Title: THE DEVELOPMENT OF THE SET OF BASIC DATA ON SPECIALISED AMBULATORY CARE (SBD-SAC) OF OSAKIDETZA. PILOT SCHEME IN THREE HOSPITALS

Authors: Yetano J, Ladrón de Guevara JM, López J, Salvador J, Rodríguez S, Ogueta M, Guajardo J.

Technology: Organization

Key words MESH: *MBDS, Artificial Intelligence, Artificial Coding, ICD-9-CM*

Date: March 2011

Pages: 58

References: 12

Language: Spanish, abstracts in English and Basque

ISBN: 978-84-457-3160-4

INTRODUCTION

Over recent years, ambulatory care has grown in importance in our hospitals. In 2009, there were 3,880,141 consultations and 912,221 hospital emergency cases in the short-term hospitals of the Basque Health Service (Osakidetza) and the care provided in day centres is increasing on a daily basis. In spite of this, there is no set of standardised data in the ambulatory area similar to the hospitalisation MBDS, which has proved to be very useful. No information is available on the illnesses of patients attended by ambulatory services nor about the procedures carried out on them. For this reason, there is no information available on the casuistics of each centre, medical service or specific doctor or on those aspects relating to the results of such care in order to be able to assess its efficiency and quality. The greatest difficulty involves the gathering of data (and especially the work of coding diagnostics and procedures) on hundreds of thousands of cases dealt with by ambulatory services.

AIMS

Define a Set of Basic Data on Specialised Ambulatory Care (SBD-SAC), develop an application that will allow the SBD-SAC to be gathered and the automatic coding of diagnostics and procedures and the implementation in three hospitals.

MATERIAL AND METHODS

The scientific literature and the regulations of a number of autonomous communities have been reviewed, a SBD-SAC with the variables to be gathered has been defined, the quality of the decentralised coding of emergency services in 4 Osakidetza hospitals has been assessed and the KodifiKa programme (automatic coding with the ICD-9-CM of the diagnostics entered by the health service professionals who provide ambulatory care) has been developed.

KodifiKa processes a literal expression of the diagnostics and is capable of coding this in accordance with the knowledge contained in a database of already coded diagnostic literals using artificial intelligence in the semantic assessment of the diagnostic expression.

Economic analysis: YES (NO) **Expert opinion:** YES (NO)

RESULTS

This paper presents an assessment of the quality of decentralised coding in emergency care episodes, the list of agreed SBD-SAC variables and data on the use of KodifiKa (80.3% of automatic coding) as well as their technical description in Annex 1.

CONCLUSIONS

The availability of a SBD-SAC in our National Health System is desirable and achievable if the Health Department normalises this, the use of the Electronic Clinical History increases and the problem of coding the diagnostics and procedures of huge amounts of cases of ambulatory care with semiautomatic systems that guarantee quality is resolved. We argue in favour of the implementation of a mixed coding programme (decentralised-centralised) such as KodifiKa.

1. INTRODUCCIÓN

El Conjunto Mínimo de Datos Básicos (CMBD) es la base de datos hospitalarios más importante que tenemos en España. Contiene datos administrativos y clínicos de los pacientes atendidos en el área de hospitalización y dados de alta. Se implantó a partir de la decisión del Consejo Interterritorial del Sistema Nacional de Salud de 1987. En la Comunidad Autónoma del País Vasco se reguló el CMBD en 1992 [1]. Supuso un avance extraordinario en los sistemas de información hospitalarios y ha permitido a lo largo de las dos últimas décadas poder conocer la casuística hospitalaria y medir aspectos cuantitativos y cualitativos de la asistencia prestada a los pacientes hospitalizados. La explotación más común del CMBD es a través de la utilización de sistemas de clasificación de pacientes como los Grupos Relacionados por el Diagnóstico (GRD) [2, 3] o para obtener información clínica [4].

Dentro de la asistencia hospitalaria, la prestada a los pacientes ingresados ha sido tradicionalmente la de mayor peso en la preocupación de clínicos y gestores debido a su complejidad y a la enorme cantidad de recursos que emplea. De ahí que los sistemas de información estén más desarrollados en esta área. Sin embargo, en las últimas dos décadas, la asistencia hospitalaria a los pacientes ambulatorios es progresivamente más relevante en nuestros hospitales pues muchos de los procesos tratados hasta ahora en régimen de hospitalización se están resolviendo de forma ambulatoria. A pesar de ello, no se dispone de un conjunto de datos normalizado en el área ambulatoria similar al CMBD de hospitalización. Esta laguna de nuestros sistemas de información está produciendo que clínicos y gestores carezcan de datos normalizados de la actividad y calidad asistencial de los procesos ambulatorios que permita su comparación y evaluación. Otro de los inconvenientes es la imposibilidad de utilización de sistemas de clasificación de pacientes ambulatorios como los Ambulatory Patient Groups (APG) [5].

Para dar una idea de la magnitud de este déficit de información (*terra incógnita* de los sistemas de información hospitalarios) es bueno cuantificarlo. Aunque sólo sea en dos de las prestaciones ambulatorias hospitalarias, en el Sistema Nacional de Salud español se produjeron en 2008 [11]:

Consultas de asistencia especializada	79.614.279
Urgencias hospitalarias	26.249.125

En asistencia especializada de Osakidetza/Servicio Vasco de Salud, durante el año 2009 y sólo referido a Consultas Externas se realizaron las siguientes:

Consultas primeras	1.133.080
Consultas sucesivas	2.746.061
Consultas totales	3.880.141

De esta enorme cantidad de asistencia prestada se tiene información del número, de su distribución por centros y servicios médicos e incluso de las demoras producidas pero faltan los aspectos clínicos. No se tiene información de qué enfermedades tenían esos enfermos atendidos ni qué procedimientos se les realizaron en las consultas. Por ello, no se tiene conocimiento de la casuística de cada centro o servicio médico o médico concreto ni de los aspectos relacionados con los resultados de esa asistencia para poder evaluar su eficiencia o su calidad.

A lo largo de la última década, a pesar de ser demandado por todos los gestores del Sistema Nacional de Salud y de algunos intentos infructuosos, ni el Consejo Interterritorial ni el Ministerio de Sanidad han sido capaces de consensuar un Conjunto de Datos Básicos de la Asistencia Ambulatoria Especializada (CDB-AAE). Aunque existen experiencias parciales [6], no las hay generalizadas a todo el ámbito ambulatorio en ninguna Comunidad Autónoma. En el País Vasco, Comunidad Autónoma que ha sido pionera en iniciativas legales relacionadas con la historia clínica [7, 8], en la utilización de sistemas de clasificación de pacientes [9] y en aplicaciones informáticas relacionadas con la documentación clínica [10], tampoco hemos sido capaces de avanzar en un CDB-AAE.

Las causas que justifican este retraso son múltiples:

1. La falta de consenso a la hora de definir el ámbito de lo que es asistencia ambulatoria especializada.
2. Las dudas respecto a si la unidad de registro debe ser el episodio asistencial o el contacto.
3. Las diferencias de criterio para la elección de la mejor clasificación con la que codificar los diagnósticos y los procedimientos ambulatorios.
4. La dificultad de recogida de los datos y especialmente la de codificar los diagnósticos y los procedimientos de cientos de miles de asistencias ambulatorias. Esta dificultad es debida a:
    - La pobre implementación de la Historia Clínica Electrónica en el área ambulatoria de la asistencia especializada.
    - La dificultad de conseguir una codificación descentralizada de calidad pues obliga a una colaboración de la totalidad de los médicos y que tengan formación en codificación.
    - La imposibilidad de hacer una codificación centralizada manual (que unos codificadores expertos codifiquen la totalidad de las asistencias ambulatorias en base a la información obtenida de la Historia Clínica Electrónica). Un hospital de referencia necesitaría decenas de codificadores para codificar manualmente su actividad ambulatoria con el consiguiente gasto.

Sin embargo, en el último lustro la situación ha mejorado y actualmente:

1. Se ha definido que el ámbito del CDB-AAE debe ser toda asistencia especializada no recogida en el CMBD de hospitalización.
2. Está claro que la unidad de información debe ser el contacto. Por ejemplo, en consultas externas se debe recoger el CDB-AAE de cada visita.
3. La clasificación más actualizada es la Clasificación Internacional de Enfermedades 9.ª, Modificación Clínica (CIE-9-MC) sin perjuicio que en el futuro el Sistema Nacional de Salud decida sustituirla por la CIE-10.
4. Están cada vez más asentados los procedimientos de registro en los sistemas informáticos y se está implementando la Historia Clínica Electrónica en Consultas Externas.
5. Se puede hacer una codificación descentralizada con apoyos informáticos que mejore la calidad o se puede hacer una codificación centralizada con codificación semiautomática.

Por todo ello, sería conveniente dar pasos hacia adelante en el registro de la actividad ambulatoria de los hospitales de la Comunidad Autónoma del País Vasco mediante la participación:

- Del Departamento de Sanidad. Consensuando y legislando mediante un decreto un CDB-AAE que promueva el registro del mismo de manera similar a lo ocurrido hace 20 años con el CMBD de hospitalización.
- De Osakidetza. Implementando sistemas de recogida del CDB-AAE y, en especial, de la codificación de Diagnósticos y Procedimientos que puedan asumir la codificación de millones de asistencias ambulatorias anuales.

2. OBJETIVOS

1.º Definir el CDB-AAE precisando el ámbito, la unidad de registro, las variables a recoger y la clasificación de enfermedades y procedimientos a utilizar.

2.º Desarrollar una aplicación que, dentro de Osabide, permita la recogida del CDB-AAE y, especialmente, la codificación automática a partir de expresiones literales de diagnósticos y procedimientos.

3.º Implementación del CDB-AAE en tres hospitales de agudos de Osakidetza.

4.º Difundir los resultados para su extensión al resto de los hospitales de Osakidetza y para su conocimiento por el resto del Sistema Nacional de Salud.

3. MATERIAL Y MÉTODOS

Se creó un grupo de trabajo con los miembros del equipo investigador. Son expertos en el mundo de la documentación clínica hospitalaria, el manejo de bases de datos y su explotación para la gestión sanitaria. Además, varios de ellos, al ser responsables directos de la gestión de varios hospitales, podían favorecer la implantación del CDB-AAE. Además de los miembros del equipo investigador se contó con informáticos con responsabilidades directivas en Osakidetza.

Los pasos que se llevaron a cabo fueron:

1.º **Revisión de la literatura** científica para valorar las experiencias de otros hospitales u otros servicios regionales de salud sobre el tema. Los artículos de experiencias en nuestro país son prácticamente inexistentes. Sí existe normativa de algunas comunidades autónomas respecto a la recogida del CMBD de algún tipo de asistencia ambulatoria como es el hospital de día o urgencias. Destacan el CMBD de Hospital de día médico y de Hospital de día quirúrgico del la Comunidad Autónoma de Andalucía [6] o el de urgencias en la de Cataluña [12].

2.º **Definición del CDB-AAE.** Tras consultar las experiencias previas y a través del consenso entre los expertos del grupo investigador se llegaron a los acuerdos siguientes:

- Ámbito de aplicación. El CDB-AAE es un registro de los datos básicos administrativos y clínicos que resume la información de cada uno de los actos asistenciales que proporciona la asistencia especializada en régimen ambulatorio, es decir, excluyendo la hospitalización tradicional. Se debe acotar el ámbito con la descripción pormenorizada de lo que incluye y de lo que excluye por normativa del Departamento de Sanidad del Gobierno Vasco.
- Unidad de registro: Es el contacto del paciente con la asistencia especializada para consulta, diagnóstico o tratamiento realizado de forma ambulatoria (sin ingresar en la hospitalización tradicional). En el caso de una consulta, la unidad de registro no es el proceso o el episodio sino cada visita.

Una vez definido el CDB-AAE se presentó junto a este proyecto al Director de Planificación y Ordenación Sanitaria del Departamento de Sanidad en Septiembre de 2009 entregándole las variables a recoger consensuadas por el grupo. Se le instó a la creación del CDB-AAE de la Comunidad Autónoma del País Vasco y a su publicación mediante un Decreto. Producto de dicha reunión fue la creación de una Comisión para el desarrollo del CDB-AAE que empezó a reunirse el 12 de noviembre del 2009.

3.º **Evaluación de la calidad de la codificación descentralizada de urgencias**. Tres miembros del equipo investigador con experiencia en codificación con la CIE-9-MC analizaron la calidad de la codificación de los diagnósticos de las urgencias atendidas en un trimestre de 2008 en los hospitales de Cruces, Basurto, Donostia y Galdakao. Se trataba de lo que los médicos de urgencia codificaron en el programa PCH de urgencias respecto a los episodios de urgencia atendidos por ellos.

4.º **Desarrollo del software de registro y salida del CDB-AAE**. Se comunicó a los responsables de la subdirección de informática de la organización central de Osakidetza que Osabide debe recoger todas las variables del CDB-AAE y especialmente la codificación de los diagnósticos generados y los procedimientos realizados en las asistencias ambulatorias. Para ello es necesario que en Osabide Global (la Historia Clínica Electrónica que está actualmente en proceso de implementación) se encaje el programa de codificación automática que surja del trabajo investigador de esta beca comisionada (KodifiKa).

En concreto, se hicieron varias reuniones con representantes de las empresas informáticas EJIE, Ibermática y Bilbomática para:

- La adaptación de los programas actuales de Osakidetza para permitir recoger el CDB-AAE, especialmente la codificación de los diagnósticos y procedimientos.

- Crear la salida del CDB-AAE.
- Desarrollo del programa KodifiKa de codificación automática con la CIE-9-MC de los diagnósticos introducidos por el profesional que presta la asistencia ambulatoria. Tras valorar varias opciones se contrató a la empresa Ibermática para el desarrollo de KodifiKa por su actitud entusiasta hacia este proyecto innovador y por su experiencia en el desarrollo de Osabide Global (la Historia Clínica Electrónica de Osakidetza) donde se tiene que encajar. Se ha encargado de su desarrollo el Instituto Ibermática Innovación.

  Se trata de un «motor» que a partir de una expresión literal de un diagnóstico es capaz de codificarlo en función del conocimiento contenido en una base de datos de literales de diagnósticos ya codificados utilizando inteligencia artificial en la valoración semántica de la expresión diagnóstica.

  KodifiKa utiliza el conocimiento de una base de datos con más de 100.000 literales de diagnósticos ya codificados con la CIE-9-MC que poseía el Hospital de Galdakao. Dicha base de datos se fue acumulando durante 2006 a 2008 a partir del programa CodeHelp de codificación de informes de alta hospitalaria. Para las pruebas de KodifiKa durante 2009 y para comenzar a trabajar en real desde Abril de 2010 utilizó dicha base de datos de partida pero, a partir de esta fecha, se alimenta del conocimiento de sus propios casos.

4. RESULTADOS

La evaluación de la calidad de la codificación descentralizada de las urgencias (programa PCH de urgencias) en cuatro hospitales Osakidetza demostró un grado variable de cumplimentación. Osciló del 52% en el Hospital de Galdakao al 93% en el Hospital de Basurto. Aunque la cumplimentación es encomiable, la calidad se juzgó deficiente pues existían graves errores:

- Codificación insuficiente (códigos de 4 cifras cuando deberían ser 5).
- Codificación incorrecta favorecida porque, en su día, se cometió el error de dar la posibilidad de que un médico pueda modificar los literales de la CIE-9-MC (que, a veces, son macarrónicos o alejados del lenguaje médico habitual) para hacerlos más «amigables». Con ello se consiguió que, por ejemplo, el código 009.3 que corresponde al literal «Diarrea presumiblemente infecciosa» cualquier médico podía cambiarlo por otro literal, «Diarrea», por ejemplo, cuando «Diarrea» se codifica 787.91. Por ello, aunque al médico no se le dejaba crear códigos nuevos y sólo podía codificarse con los oficiales de la CIE-9-MC, al dejarle cambiar los literales, se consiguió una CIE-9-MC errónea y el médico al querer codificar «Diarrea», buscaba por literal y le aparecía un 009.3 y lo elegía cuando lo correcto era que hubiese codificado 787.91.

La lista de variables del CDB-AAE consensuada y entregada al Director de Planificación y Ordenación Sanitaria del Departamento de Sanidad en Septiembre de 2009 es:

1. Tipo de asistencia:
   - Urgencia hospitalaria.
   - 1.ª visita en Consulta Externa.
   - Visita sucesiva en Consulta Externa (cada contacto-visita).
   - Hospital de Día (cada contacto).
   - Hospitalización a domicilio (cada episodio).
   - Otros tipos de asistencia ambulatoria especializada.
2. Número de episodio o de la identificación de la consulta-visita.
3. Número de Historia Clínica del hospital que se trate (o n.º de episodio de urgencias).
4. CIC (Código de identificación corporativa de Osakidetza) si tiene.
5. Identificación del centro sanitario (hospital o ambulatorio).
6. Fecha de nacimiento del paciente.
7. Sexo del paciente.
8. Domicilio, Municipio y Provincia del paciente.
9. Garante.
10. Procedencia. Si es de otro hospital se identificará con el código del Catálogo Nacional de Hospitales.
11. Fecha de ingreso o de comienzo del contacto.
12. Hora de ingreso o de comienzo del contacto.
13. Fecha de alta o de finalización del contacto.
14. Hora de alta o de finalización del contacto.
15. Duración del contacto en horas (calculado a partir de 14 del 13 menos 12 del 11).
16. Destino. Si es a otro hospital se identifica con el código del Catálogo Nacional de Hospitales.
17. Diagnóstico motivo del contacto. Codificado mediante la CIE-9-MC vigente o la que en el futuro pueda declarar oficial le Ministerio de Sanidad.
18. Otros diagnósticos. Codificado mediante la CIE-9-MC vigente o la que en el futuro pueda declarar oficial le Ministerio de Sanidad.
19. Procedimientos. Codificado mediante la CIE-9-MC vigente o la que en el futuro pueda declarar oficial le Ministerio de Sanidad.
20. Código de la sección de urgencias que da el alta (en caso de asistencia en Urgencias).

Se desarrolló el programa Kodifica (ver su descripción en el Anexo 1) y se implantó en dos hospitales.

El desarrollo de KodifiKa se comenzó en 2008 y se consiguió un programa de codificación que analiza una expresión diagnóstica introduci-

da por el médico, la trata semánticamente y la compara con la base de datos CodeHelp (de más de 100.000 expresiones diagnósticas ya codificadas) y la clasifica de 4 maneras:

1. *Codificándola de forma segura* pues la encuentra exacta en la Base de Datos de Codehelp ya codificada por un codificador experto. «Úlcera de pierna derecha», por ejemplo, lo encuentra exacto en la base de datos ya codificada por lo que le asigna un código con seguridad.
2. *Codificándola de forma altamente probable* pues, aunque no la encuentra exacta en la Base de Datos de CodeHelp, por búsqueda semántica es capaz de asignarle un código muy probable. «Úlcera en pierna derecha» que, aunque no la encuentra exacta, es capaz de codificarla a partir de otras expresiones parecidas valorando los pesos de las diferentes palabras y la frecuencia histórica.
3. *No sabe codificarlo*. La expresión diagnóstica introducida por el médico no se encuentra exacta en la Base de Datos de Codehelp ni se encuentran otras expresiones, que no siendo exactas, permitan a KodifiKa mediante análisis semántico asignarle un código CIE-9-MC con alta probabilidad de acierto. En este caso el diagnóstico queda pendiente de codificación por un experto pero, una vez realizada, ese conocimiento pasa a la base de datos del programa para su utilización en el futuro.
4. *No es un diagnóstico codificable*. Por ejemplo, si el médico introdujo «No acude» y el programa tiene conocimiento para saber que esa frase no es un diagnóstico codificable.

Lo que se ha diseñado es que KodifiKa sepa codificar automáticamente y con calidad (apartados 1, 2 y 4 del párrafo anterior) un porcentaje muy elevado de los diagnósticos. Los casos del apartado 3 deben ser codificados por un experto manualmente pero KodifiKa le propone una codificación probable facilitándole la tarea. Los casos del apartado 3, una vez codificados por el codificador, pasan a la base de Datos de literales codificados para su utilización en el futuro.

Kodifika se encajó en Osabide Global (Historia Clínica Electrónica) a comienzos de 2010 para poder trabajar sobre los diagnósticos introducidos por los médicos en las evoluciones realizadas en dicha Historia Clínica Electrónica. Gracias a que en Abril de 2010 se comenzó la implementación de Osabide Global se ha podido probar KodifiKa en real.

Los centros y servicios en los que se usa desde abril de 2010 son:

- Hospital de Cruces en todas las evoluciones de Hospitalización a Domicilio.
- Hospital de Txagorritxu en las consultas externas de los servicios:
  - Digestivo.
  - Dermatología.
  - Oncología médica.
  - Unidad del Dolor (de Anestesia).
  - Respiratorio.
  - Oftalmología.
  - Traumatología.
- Ambulatorio de Olaguibel en Vitoria:
  - Oftalmología.
  - Traumatología.
- Ambulatorio de Lakuabizkarra en Vitoria:
  - Dermatología.

Resultados de KodifiKa desde Abril de 2010 aplicado a los hospitales de Txagorritxu (consultas externas de 7 servicios) y Cruces (visitas de Hospitalización a domicilio).

a) N.º de Consultas y visitas realizadas en Osabide Global 3.187

b) N.º de Consultas y visitas de *a)* tratadas por Kodifika 2.185*

c) N.º de diagnósticos de las 2.185 Consultas y visitas de *b)*..... 2.401

* En las 3.187 consultas o visitas se formuló un diagnóstico en todas pero sólo entraron en KodifiKa las 2.185 en las que se formuló por primera vez o se cambió o se añadió un nuevo diagnóstico.

Los Diagnósticos codificados automáticamente por KodifiKa fueron el 80,3%.

5. CONCLUSIONES

Tener un CBD-AAE en nuestro Sistema Nacional de Salud es algo deseable y alcanzable a pesar de la dificultad de la recogida del elevado número de asistencias ambulatorias.

Se ha llegado a la conclusión de que la codificación descentralizada por parte del médico es de mala calidad por:

- Falta de formación en codificación.
- Falta de tiempo del profesional para hacer una codificación meticulosa.
- Rechazo de la terminología de la CIE-9-MC que, a veces, se aleja de la utilizada comúnmente por los médicos.

Por todo ello, defendemos la implementación de un programa de codificación mixto (descentralizado-centralizado) como KodifiKa. Es descentralizado en cuanto a que es el médico el que introduce la expresión diagnóstica en lenguaje natural libre (en la evolución de la Historia Clínica Electrónica). La formulación del diagnóstico es completamente libre pero a partir de la introducción de la 5.ª letra le aparece un desplegable de sus diagnósticos anteriormente formulados por él permitiéndole seguir escribiendo libremente o seleccionar una de las expresiones que le aparecen. Si se elige un diagnóstico usado anteriormente del desplegable la codificación es segura pues ya está codificado de días anteriores. El médico no codifica, no elige un código, sino que formula un diagnóstico libremente en lenguaje natural con lo que se consigue mayor riqueza expresiva. Es un sistema de codificación centralizado, es decir, codificado por expertos de la Unidad de Codificación que garantiza la calidad de la codificación pero automático en un porcentaje mayor del 80% (ya codificado por un codificador o basado en lo codificado por un codificador en el pasado) y manual en un porcentaje menor del 20% (pero con propuestas por parte del programa que aligeran el trabajo del codificador). Sin duda, en el futuro inmediato se podrá subir el porcentaje de codificación segura y de calidad a más del 90% si se emplean técnicas de inteligencia artificial que valoren no sólo la expresión diagnóstica sino todo el contexto, es decir, todo el texto de la evolución de la consulta o datos tan sencillos de abordar en la Historia Clínica Electrónica como la edad o el sexo del paciente.

Con este estudio se ha alcanzado, al menos parcialmente, los tres primeros objetivos: consensuar un CDB-AAE en Osakidetza (pendiente de culminar por parte de la Comisión del Departamento de Sanidad para toda la Comunidad), desarrollar un programa de codificación semiautomático de diagnósticos (KodifiKa) y su implementación dos hospitales de agudos de Osakidetza. No se han conseguido la salida del CDB-AAE normalizado de los sistemas de información de los hospitales de Osakidetza ni la divulgación de este trabajo pero son objetivos alcanzables en 2011.

Como limitación de este estudio hay que destacar la debilidad de los datos presentados pues el periodo de utilización del programa KodifiKa (desde abril de 2010) es demasiado corto.

6. RECOMENDACIONES

Para conseguir a corto plazo el CDB-AAE en la Comunidad Autónoma del País Vasco es necesario:

- Que el Departamento de Sanidad normalice el CDB-AAE actualmente en estudio.
- Se extienda a todos los niveles la implementación de la Historia Clínica Electrónica en asistencia ambulatoria especializada.
- Se potencien los programas de codificación semiautomática de los diagnósticos y procedimientos que aseguren la calidad.

BIBLIOGRAFÍA

1. Departamento de sanidad. Decreto 303/1992 de 3 de noviembre de 1992. BOPV, n.º 234, 1 de diciembre 1992.
2. Casas M. Los grupos relacionados con el diagnóstico: experiencias y perspectivas de utilización. Barcelona, Masson SA, 1991.
3. AP-GRDs. All Patient Diagnosis Related Groups. Definition Manual. Version 21.0. 3M Health Information Systems. 2003.
4. García de Jalón J, Nuín MA, Panizo A. Utilidad del conjunto mínimo básico de datos (CMBD) en la vigilancia de las infecciones nosocomiales. Anales del sistema sanitario de Navarra 2000; 23(2): 237-246.
5. Ruiz de la Prada Mac-Crohon L. Sistema de clasificación de pacientes ambulatorios (APGs).Todo hospital 1997: 139:19-21.
6. Servicio andaluz de la salud. Manual de instrucciones del CMBD: hospitalización, hospital de día quirúrgico y hospital de día médico. Sevilla 2006.
7. Decreto 272/1986, de 25 de noviembre, Uso de la Historia Clínica en los centros hospitalarios de la Comunidad Autónoma del País Vasco.
8. Decreto 45/1998, de 17 de marzo, por el que se establece el contenido y se regula la valoración, conservación y expurgo de los documentos de las Historias Clínicas hospitalarias.
9. Manual de descripción de los Grupos Relacionados con el Diagnóstico. Osakidetza/ Servicio Vasco de Salud. Noviembre 2000.
10. Yetano Laguna J, López Arbeloa G, Guajardo Remacha J, Barriola Lerchundi MT. Kliniker. Un sistema de información para clínicos y gestores. VIII Jornadas de gestión y evaluación de costes sanitarios. Salamanca, 7,8 y 9 de Junio de 2006.
11. Estadística de establecimientos sanitarios con régimen de internado. 2008. Instituto de Información Sanitaria. Ministerio de Sanidad y Política Social.
12. Registre del conjunt mínim bàsic de dades dels serveis d'urgències de Catalunya. Manual de notificació. Servei Catalá de la Salut. Prova pilot. Març 2009.

ANEXOS

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Título de Proyecto:
KodifiKa

Lista de participantes (entidad y rol – líder o participante):
Osakidetza (líder), Instituto Ibermática Innovación (participante)

Entregable E1.1

«KodifiKa. Modelo conceptual y arquitectura»

1. INTRODUCCIÓN

1.1. Resumen del objetivo del proyecto

El objetivo principal del programa «KodifiKa», es investigar la respuesta que desde las TICs (Tecnologías de la Información y la Comunicación) puede darse a las necesidades actuales y futuras en el ámbito de la gestión de grandes masas de información dispersas en entornos de bases de datos nativos y plataformas empresariales, aplicadas a la ayuda en la codificación médica.

El programa KodifiKa, es una ayuda a la codificación automática de la nomenclatura CIE-9-MC, que se corresponde con de los diagnósticos introducidos por el profesional que presta la asistencia ambulatoria. Se trata de un «motor» que a partir de una expresión literal de un diagnóstico es capaz de codificarlo en función del conocimiento de una base de datos de literales de diagnósticos ya codificados utilizando inteligencia artificial en la valoración semántica.

Tras valorar varias opciones se contrató a la empresa Ibermática para el desarrollo de KodifiKa por su actitud entusiasta hacia este proyecto innovador y por su experiencia en el desarrollo de Osabide Global, la Historia Clínica Electrónica de Osakidetza, donde se tiene que encajar. Se ha encargado de su desarrollo la Unidad de Innovación de Ibermática.

El objetivo final es, que, el nuevo concepto de sistema de gestión de búsquedas responda a las siguientes características:

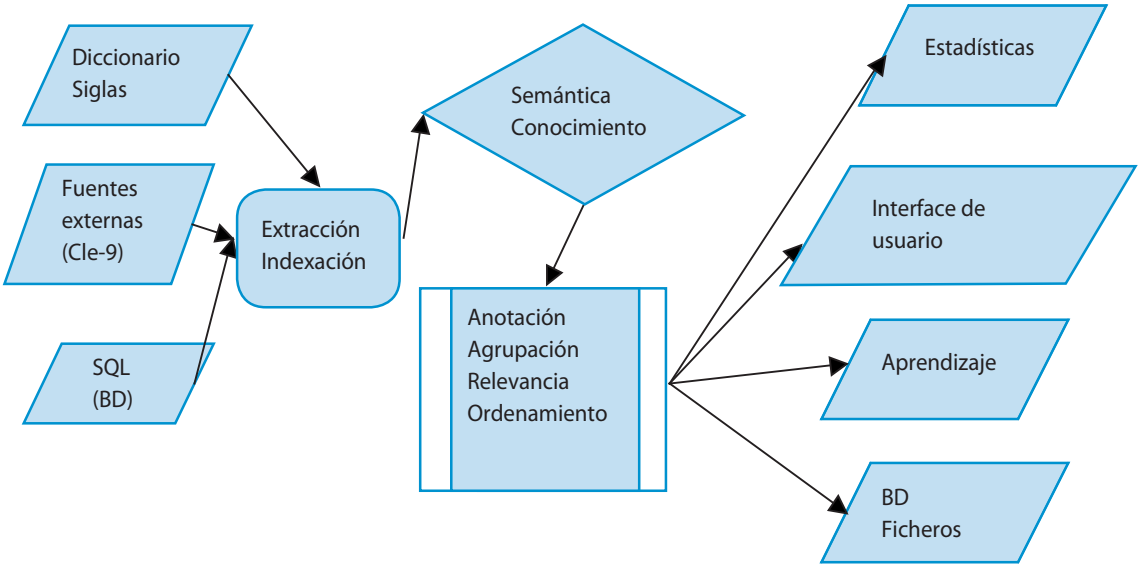
- Extendido, en el sentido de que sea a su vez plataforma de desarrollo de servicios que pueda acompañar a las empresas en sus procesos de expansión y generación de valor.
- Modular, utilizando estándares de integración entre módulos que favorezcan la integración con distintos módulos de diversos proveedores.
- Flexible, para poder adaptarse a formas de trabajo cambiantes.
- Inteligente, en el sentido poder traducir consultas en lenguaje natural a un formato de meta-datos que permita la navegación por la información correcta.
- Integrador de procesos de negocio por encima de lo que puede ser una orientación funcional.
- Con capacidad de adaptación a las personas.
- Semántico: que sea capaz de interpretar la información de entrada, no sólo de una manera morfosintáctica, sino que comprenda cuáles son los argumentos de la pregunta, para devolver respuestas inteligentes a modo de interacción en lenguaje natural.

De cara al usuario final, el interface de salida es responsabilidad de la aplicación que contiene el motor del buscador semántico, «incrustado» en los sistemas nativos como partes natural del sistema propie-

tario, y enlazándose con el motor de búsqueda semántico por medio de una conexión única a la base de datos propietaria. En definitiva, el motor semántico accederá a ciertas tablas para recoger la información de las consultas, para indexar la información tanto de los sistemas internos como de los externos, y devolverá el resultado de las consultas ordenadas en otra tabla de salida, desde dónde el interface de usuario recogerá la información y la mostrará con la apariencia natural del sistema propietario.

La siguiente figura muestra una visión general del sistema:

Figura 1. **Arquitectura general del sistema**



1.2. Objetivos del documento

El principal objetivo de este documento es describir con un grado de detalle suficiente el diseño que se ha definido para implementar las distintas funcionalidades del sistema KodifiKa y que dé respuesta a los requisitos planteados.

Se analizan por un lado el diseño del módulo de conexión con las fuentes externas e internas (robots y enlaces), el modulo de indexación, el módulo de comprensión semántica y desambiguación, el módulo de extracción de la información relevante (tratamiento de lenguaje natural), su anotación y ordenación, y por último, el módulo de generación semántica de las soluciones propuestas.

2. ENTORNO DE DESARROLLO Y EXPLOTACIÓN

El entorno de desarrollo escogido se basa en herramientas y sistemas informáticos de uso extendido en el ámbito de las tecnologías de información, por lo que no se prevé un riesgo asociado a este aspecto. En esta sección se describen las configuraciones hardware y software utilizadas en el proyecto, diferenciando los tres ámbitos siguientes:

- Entorno de desarrollo,
  - Servidor: Oracle.11g.
  - Desarrollo WPF (Windows Presentation Foundation).
  - Visual Studio 2008.

- Entorno de ejecución o despliegue (servidor) y
  - Servidor: Oracle.11g.
  - Windows Server 2008.
  - Web services de Osabide.

3. ARQUITECTURA DEL SISTEMA

En la presente sección se describe de forma detallada la arquitectura del sistema, analizando cada módulo y funcionalidad desarrollado.

3.1. Características generales

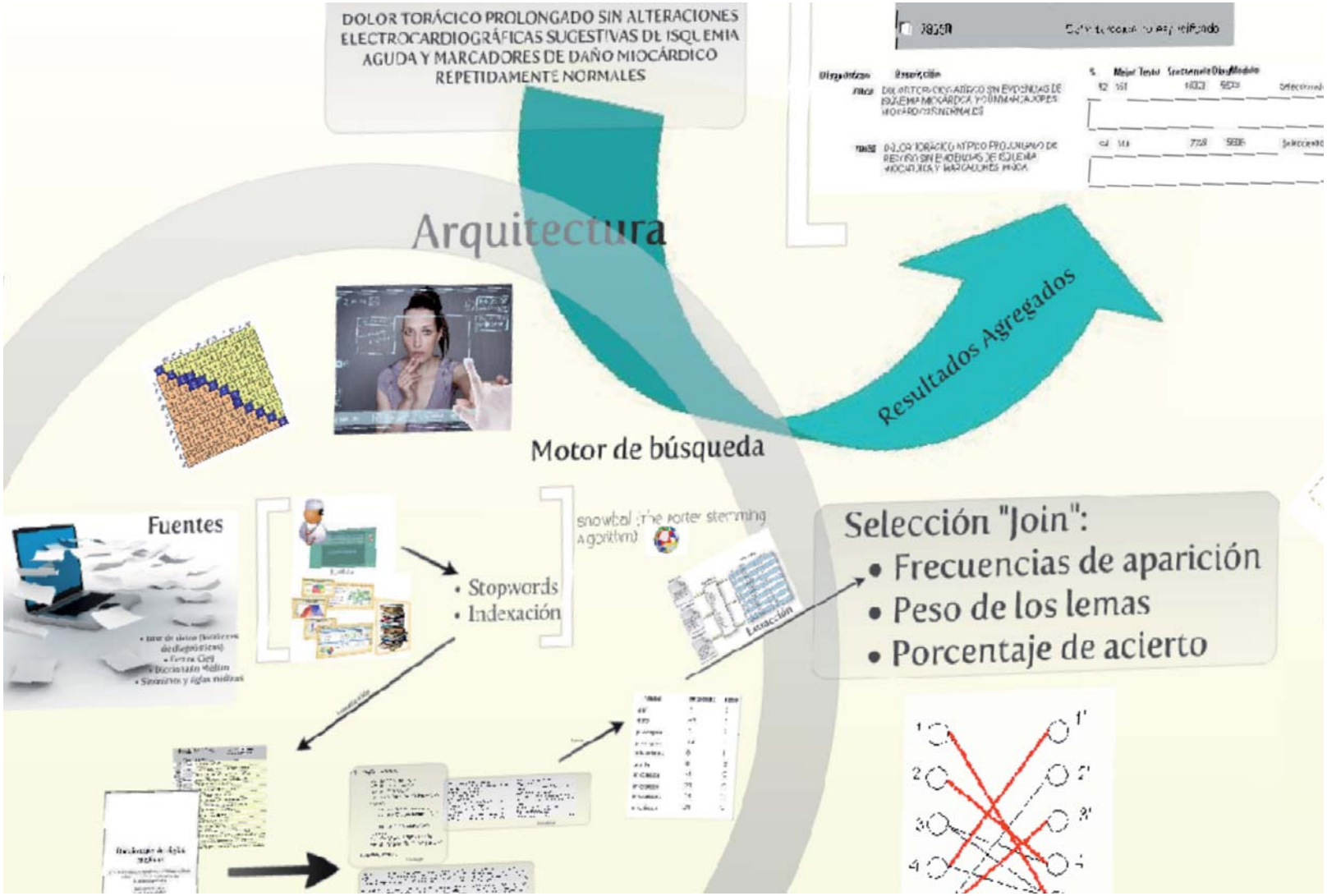
Desde un punto de vista de desarrollo y mantenimiento

- **Modularidad**. El sistema se concibe de forma totalmente modular. Las entradas y salidas de cada módulo están perfectamente identificadas. Esto permitiría en un futuro adaptar fácilmente el sistema de búsqueda para otros usos, así como implantar el servidor de datos y la lógica desarrollada con otro sistema de captación de datos distinto al utilizado en KodifiKa. Las interfaces de usuario final también se pueden adaptar y modificar de forma totalmente independiente a las características del sistema de búsqueda subyacente.
- **Escalabilidad y adaptabilidad al sistema propietario**. El diseño planteado, tanto de la aplicación como de la base de datos puede ser replicado a múltiples sistemas del mismo tipo. Además esta información se podrá actualizar de forma dinámica según la obra evolucione.
- **Facilidad de mantenimiento**. Desde el punto de vista de los puestos de los usuarios, el sistema derivado de este proyecto no implica ninguna labor de mantenimiento adicional a la de sus herramientas habituales de trabajo, ya que se trata de un cliente ligero basado en accesos a bases de datos. Atendiendo al servidor, el mantenimiento del mismo es mínimo y la base de datos y sus procedimientos están implementados en Oracle, al igual que el resto de sistemas de Osakidetza, por lo que su mantenimiento (backups, configuración, permisos…) serán similares al resto de aplicaciones ya presentes y se podrá integrar en la gestión informática general , en especial, en Osabide Global, sin otros requisitos específicos.

Desde el punto de vista del usuario final del sistema, éste presenta las siguientes características:

- **Facilidad de configuración y uso**. Las interfaces de usuario final están concebidas como una aplicación web, por lo que se trata de un entorno muy extendido en la actualidad y al que los usuarios están acostumbrados. Se ha primado un diseño sencillo, basado en elementos estándar (menús de selección, formularios, pestañas, botones, etc.), complementado con información visual (gráficas y códigos de colores, que ayudan a comprender mejor los resultados). La configuración que deba hacer el usuario final que tenga los permisos para ello será básica y a través de la propia aplicación, sin necesidad de entender los entresijos del servidor de datos subyacente.
- **Entorno estándar**. Al tratarse de una aplicación embebida en un entorno web, para el usuario es transparente el motor de búsqueda, y no será necesaria ninguna modificación a la configuración que tenga actualmente en el acceso a sus sistemas.
- **Bajo coste de implantación**. El coste de implantación es bajo, debido a que se embebe como una aplicación Web en el sistema en el que los usuarios estén trabajando.

Figura 2. **Modelo conceptual de búsqueda semántica**



3.2. Modelo conceptual

En este apartado se describe el modelo conceptual de información sobre el cual operan los distintos módulos que posteriormente se describen en el documento. La implementación física de este modelo conceptual en Oracle, con la descripción de todas las tablas y vistas junto con sus campos y procedimientos almacenados se describe en detalle en el siguiente apartado.

3.2.1. Modelo conceptual del buscador semántico

El sistema traerá una serie de datos predefinidos (tablas maestras) que no se verán modificados en el uso del sistema y que son los siguientes:

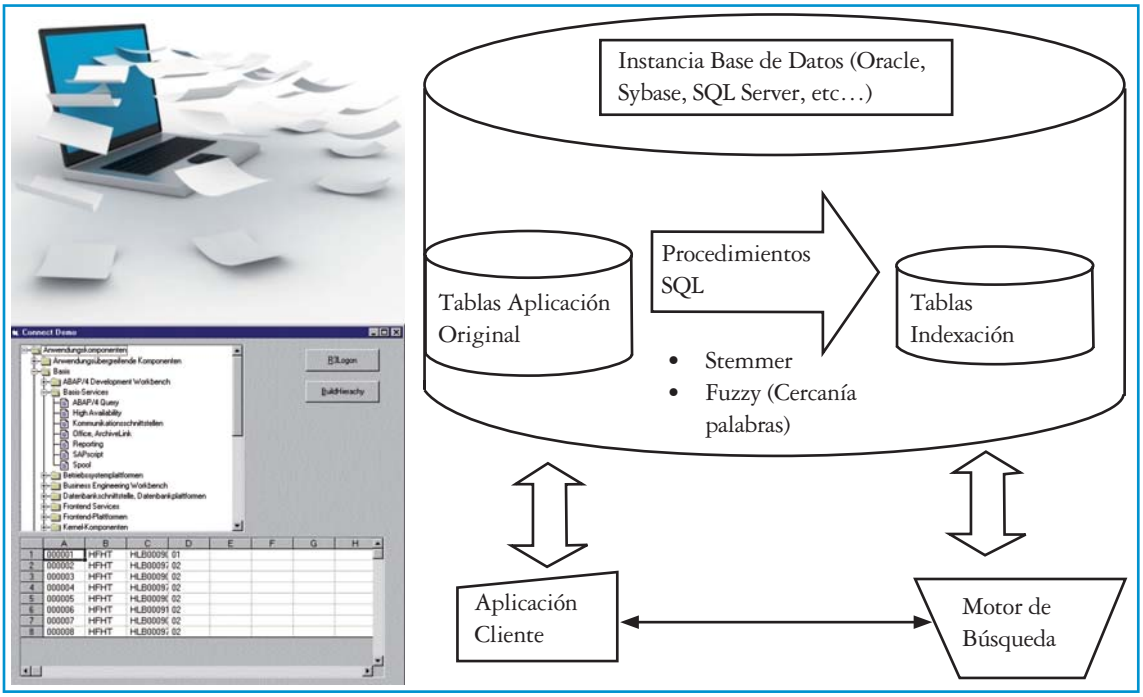
- T_real: Tabla de indexación de la información anotada.
- T_real_agrup: Tabla de agrupación de la información anotada, con información sobre la media y moda de la aparición de los términos. Necesaria para los algoritmos de cálculo de la frecuencia y la frecuencia inversa.
- Stopword: Tabla que contiene las palabras a no incluir en las anotaciones. Obtenida de Snowball (http://snowball.tartarus.org/).
- Diccionario: Tabla que contiene un diccionario en diversos idiomas, para la corrección ortográfica en función de su distancia de Levenshtein.
- Sinónimos: Tabla de sinónimos. Obtenida del tesauro Snowball.
- Temp-stem: Tabla contenedora de la consulta normalizada.
- Temp_avisos_porcentaje: Tabla final con las mejores búsquedas ordenadas por relevancia.

3.2.2. Módulo de conexión con las fuentes externas e internas

Las pruebas de integración del sistema están enfocadas a la extracción de información y su anotación sobre una base de datos (Osabide Global), con la información de CodeHelp (Osakidetza), en la que se aplicará una indexación para organizar la información relevante. Independientemente de este contexto, el sistema estará preparado para realizar conexiones y anotaciones tanto a bases de datos internas, cono a fuentes externas, de forma que el motor de conexión funcione de igual forma para ambos tipos de fuentes, y además, sea independiente de plataforma, sea está un entorno de escritorio, o un entorno Web.

- El proceso de indexación y anotación de las fuentes relevantes será un proceso batch que se realice por la noches, y se basará en una función externa al motor (**Indexa_incidencia)**, en la que se especificarán una serie de filtros en los que el usuario o el administrador del sistema podrá rellenar, de forma que se planifiquen las anotaciones. Los valores de estos filtros serán parametrizables, pero tendrán en común los siguientes valores:
  - Hora de ejecución: Hora de lanzamiento del proceso de anotación.
  - Fecha de actualización: Fecha a partir de la cual las fuentes se comienzan a anotar.
  - Tipo de Fuente.
  - Url.
  - Conexión B.D.
- Las conexiones a la base de datos se realizarán desde el propio servidor, con conexión nativa de VisualStudio a Oracle.

Figura 3. **Módulo de conexiones**



3.2.3. Modelo conceptual de indexación

Este modelo describe la forma en la que la información se anota y se indexa, de forma que se extrae la información sensible al contexto de interés, y se registra en la base de datos, permitiendo posteriores consultas.

Esta acción se realiza fundamentalmente con una función, construida de forma multiplataforma y que es capaz de ser llamada por cualquier plataforma y lenguaje de programación.

La función tiene como entrada el idioma del texto que vamos a indexar, la tabla de palabras no relevantes (stopwords), y el idioma, y devuelve una lista de palabras indexadas por su lema según el algoritmo de «stemmer» de Porter, reduciendo las miles de palabras de un documento a sus raíces (facturación → factur, facturé → factur, facturado → factur), con lo que ya tenemos un índice por el que, posteriormente, navegaremos de forma inversa para buscar la información.

Con esta información, ya tenemos implícitamente la relevancia de orden de los resultados, basada en la función de «frecuencia», asignada a un documento, que, básicamente, es proporcional a la «moda» de aparición de los términos buscados en cada documento, e inversamente proporcional a la aparición de dichos términos tienen en el conjunto de información.

$$F: \ BinIDF\left(\vec{t}_i, \vec{d}_j\right) = \begin{cases} 1+\log\left(\dfrac{N}{df\left(\vec{t}_i\right)}\right), & \text{si } f_{ij} \neq 0 \\ 0, & \text{si } f_{ij} = 0 \end{cases}$$

(Por ejemplo, en un entorno de diagnósticos, la palabra «dolor» no tendrá mucha relevancia para la indexación, ya que aparecerá en la mayoría de los documentos, es decir, no es un término que permita clasificar bien la información).

3.2.4. Modelo conceptual de comprensión semántica y desambiguación (tratamiento de lenguaje natural)

Los buscadores semánticos van un poco más allá, e intentan, dentro de un dominio concreto, como lo pueda ser en este caso los diagnósticos médicos, o gestión de incidencia en mantenimientos, seleccionar los mejores resultados en base a una búsqueda no tan lineal, sino dando un peso concreto a cada uno de los índices extraídos en el paso anterior. Para realizar esta asignación de pesos, o bien, utilizamos el conocimiento de un experto, o bien, nos basamos en estructuras ya definidas por dichos expertos, y que están accesibles en formato RDF. En ambos casos, estamos hablando de ontologías, que no dejan de ser la implantación de un conocimiento en una representación jerárquica.

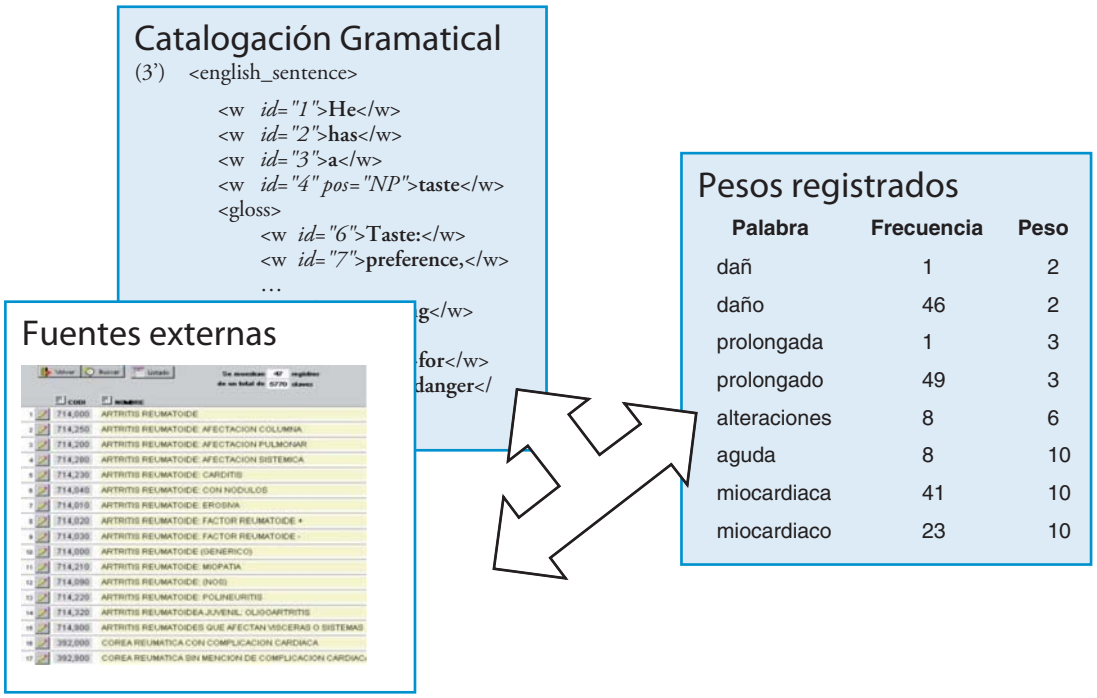
Cuando existan ontologías de un contexto determinado, podremos utilizarlas para gestionar los pesos de los lemas, y si no existen, pondremos la valoración de los pesos en la base de datos en función a la experiencia de un experto.

Pero una primera aproximación sencilla es la de utilizar un analizador gramatical que gestione la categoría gramatical de los lemas, y asignando unos pesos de forma universal, dando más «fuerza» a los sustantivos, que a los adjetivos, por ejemplo «Pedido» es más importante que «facturado», teniendo en cuenta que los documentos ya están filtrados por la selección previa de los aciertos en lemas.

En base a esta información, y con la ayuda de un «desambiguador» gramatical, es decir, un sistema que, en base a las frases introducidas (en este caso, a la información de las descripciones extraídas de las fuentes), se determina la función gramatical de cada palabra, y por ende, de cada lema. Así, en los diagnósticos médicos, habitualmente, es más importante el sujeto, que el adjetivo.

Además, a esta política de asignación de pesos, se le ha unido el conocimiento de expertos en la codificación, de forma que se ha identificado que en los textos oficiales de la codificación CIE9-MC, los primeros lemas de cada descripción tienen más peso que sobre el resto, así que se le ha asignado esta política complementaria a los pesos asignados.

Figura 4. **Política de Pesos**



Por ejemplo, en el siguiente texto escrito por un médico: «*Dolor torácico prolongado sin alteraciones electrocardiográficas sugestivas de isquemia aguda y marcadores de daño miocárdico repetidamente normales*», en una primera criba, con las palabras en el texto, ya se realiza un filtrado de todas la información de la base de datos. Sin embargo, dentro de todos los posibles diagnósticos que pueden tener relación con este texto (en concreto, 22 diagnósticos), tenemos que seleccionar aquellos que mejor se «adecuen» a lo que, semánticamente, quiere decir el médico. Para ello, no tiene el mismo peso las palabras: normales (peso=30), isquemia (peso=30), torácico (peso=10), miocárdico (peso=10), es decir, isquemia tiene más peso que los adjetivos torácico y miocárdico, pero normal tiene más peso, puesto que puede ser que exista un «marcador anormal».

Estos pesos permiten romper la «ambiguedad» semántica de los resultados, de forma, que de los 22 posibles códigos a seleccionar, (muestro los tres primeros en la imagen):

Figura 5. **Ejemplo de resolución de ambigüedades**

Diagnósticos	Descripción	%	Mejor Texto	Frecuencia	DiagMódulo
78651	DOLOR TORÁCICO ATÍPICO SIN EVIDENCIAS DE ISQUEMIA MIOCÁRDICA Y CON MARCADORES MIOCÁRDICOS NORMALES	92	151	963	6523
78650	DOLOR TORÁCICO ATÍPICO PROLONGADO DE REPOSO SIN EVIDENCIAS DE ISQUEMIA MIOCÁRDICA Y MARCADORES MIOCA	64	114	7228	15606
78659	DOLOR ATÍPICO OSTEOMUSCULAR PROLONGADO CON MARCADORES DE DAÑO MIOCÁRDICO NORMALES Y SIN EVIDENCIAS D	85	106	1047	41083

Ordenados en función de la moda, la frecuencia inversa, y los pesos semánticos de los lemas, se selecciona al respuesta correcta (78651) a la búsqueda realizada.

Para la realización del cálculo de la función gramatical, se ha utilizado la plataforma **Treetagger**, con enlaces al corpus español CRATER, aplicada sobre los lemas obtenidos.

El sistema se complementa con una gestión de sinónimos, en base a los distintos contextos en los que nos movamos (industrial, médico, etc…).

3.2.5. Modelo conceptual del módulo de extracción de la información relevante

Una vez extraído el conocimiento, este hay que ordenarlo por importancia, y por grado de satisfacción a la pregunta del usuario. Para ello, se utiliza, tal y como se ha explicado anteriormente, los siguientes conceptos:

- **Porcentaje**: Número de concordancias en lemas entre la pregunta y las posibles respuestas. Si es un 100%, quiere decir que todos los lemas presentados en la pregunta están en la respuesta, y si hay más de una repsuesta que mantienen un 100%, habrá que priorizarlas, con dos conceptos más, la frecuencia y la frecuencia contextual.

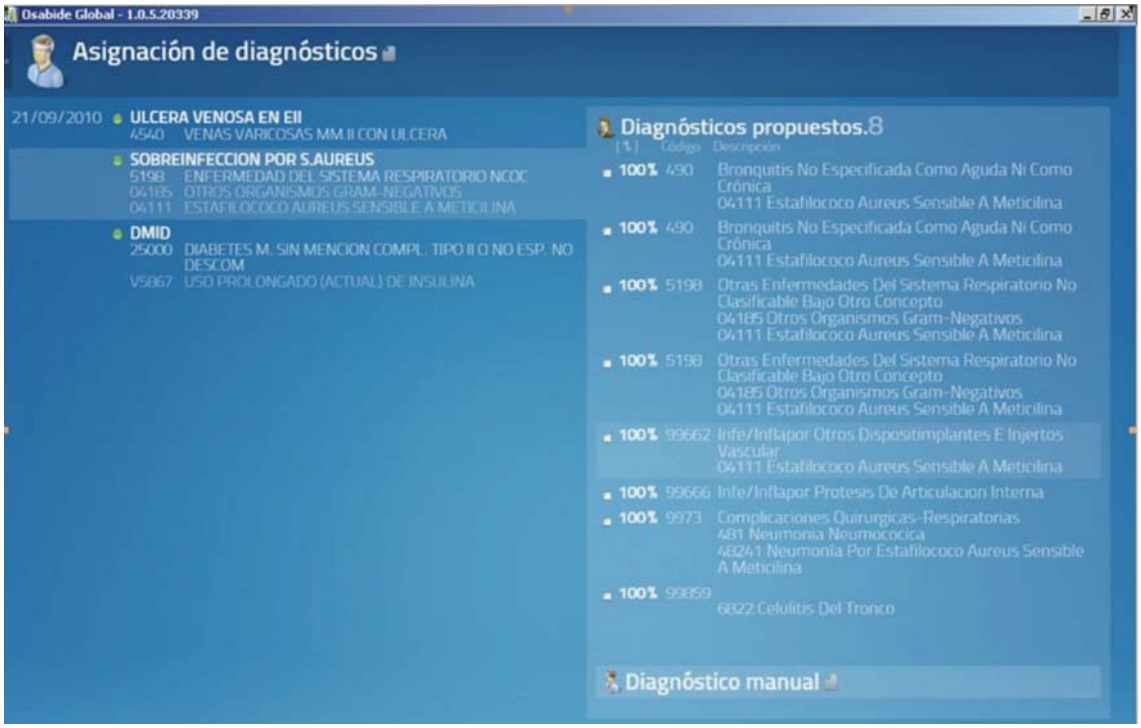
- **Frecuencia**: Une dos conceptos, la frecuencia que cada lema aparece en documento (cuantas más veces aparezca en un documento un lema buscado, más importancia tendrá dicho documento), pero a su vez, el número de veces que dicho lema aparece en el resto de documentos, es decir, si un lema aparece mucho en un documento, pero también aparece mucho en el resto, dicho lema no es representativo para discriminar información «sensible» entre el conjunto de datos.
- **Frecuencia contextual**: Dentro de un mismo documento, pueden aparece lemas buscados, pero en distintos puntos del documento. La frecuencia textual indica lo cercanos que los lemas buscados están en la respuesta, y además, indica lo cercano que está el texto a la formulación realizada en lenguaje natural en la pregunta. Difiere en el porcentaje en que, el porcentaje en un valor de inclusión, mide si los lemas incluidos en la preguntan están o no en la respuesta, y en qué grado, pero no miden cuán es la cercanía de las respuesta a la pregunta, es decir, cuantos lemas que están en la respuesta, no están en la pregunta. Además, la frecuencia contextual tiene en cuenta la frecuencia de los términos en el documento que coinciden con la pregunta, así que dos documentos nunca podrán tener la misma frecuencia contextual con respecto a una pregunta.

Frecuencia Contextual = (frecuencia_lema * num_lemas_pregunta) / num_lemas_respuesta

El orden de aparición de los resultados está presentado en orden descendente por:

1. Frecuencia contextual.
2. Porcentaje.
3. Frecuencia.

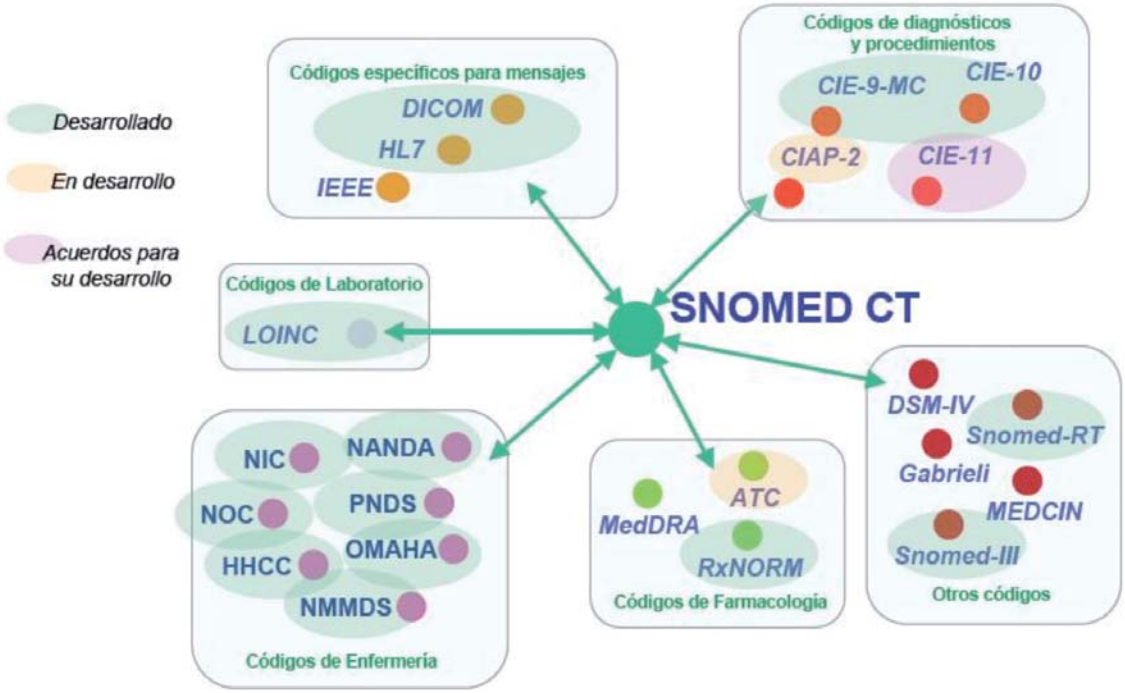
Figura 6. **Interface de usuario**



3.3. Próximos pasos

Por un lado, en futuros trabajos, se pretende incorporar la información contextual de los evolutivos a la ayuda a la codificación, de forma que datos como el género del paciente, patologías anteriores y otra información no estructurada nos pueda ayudar a determinar la «conciencia situacional» del paciente y ajustar más los resultados resolviendo la «ambigüedad» de una forma más precisa.

Figura 7. **Próximos pasos: Inclusión de ontologías en KodifiKa**



Por otro lado, en la relación del tratamiento del lenguaje natural, y su enlace con las posibles ontologías contextuales, si las hay, que puedan servir como punto de partida para la comprensión del texto, tal y como se detalla en el estado del arte, se utilizará la herramienta GATE, que nos suministra funcionalidad en dos vertientes:

- Gestión de ontologías propias del lenguaje, en castellano, basado en CRATER.
- Gestión de ontologías externas contextuales. Por ejemplo, en un entorno médico, podríamos incorporar a GATE la ontología SNOMED, que nos serviría para ubicar en clases las categorías gramaticales extraídas anteriormente, y además, permitir anotar en lenguaje médico los documentos relevantes.

Esta funcionalidad podrá enriquecer la información mete-lingüística de la búsqueda, añadiendo sinónimos, antónimos y mayor expresividad a la consulta inicial.

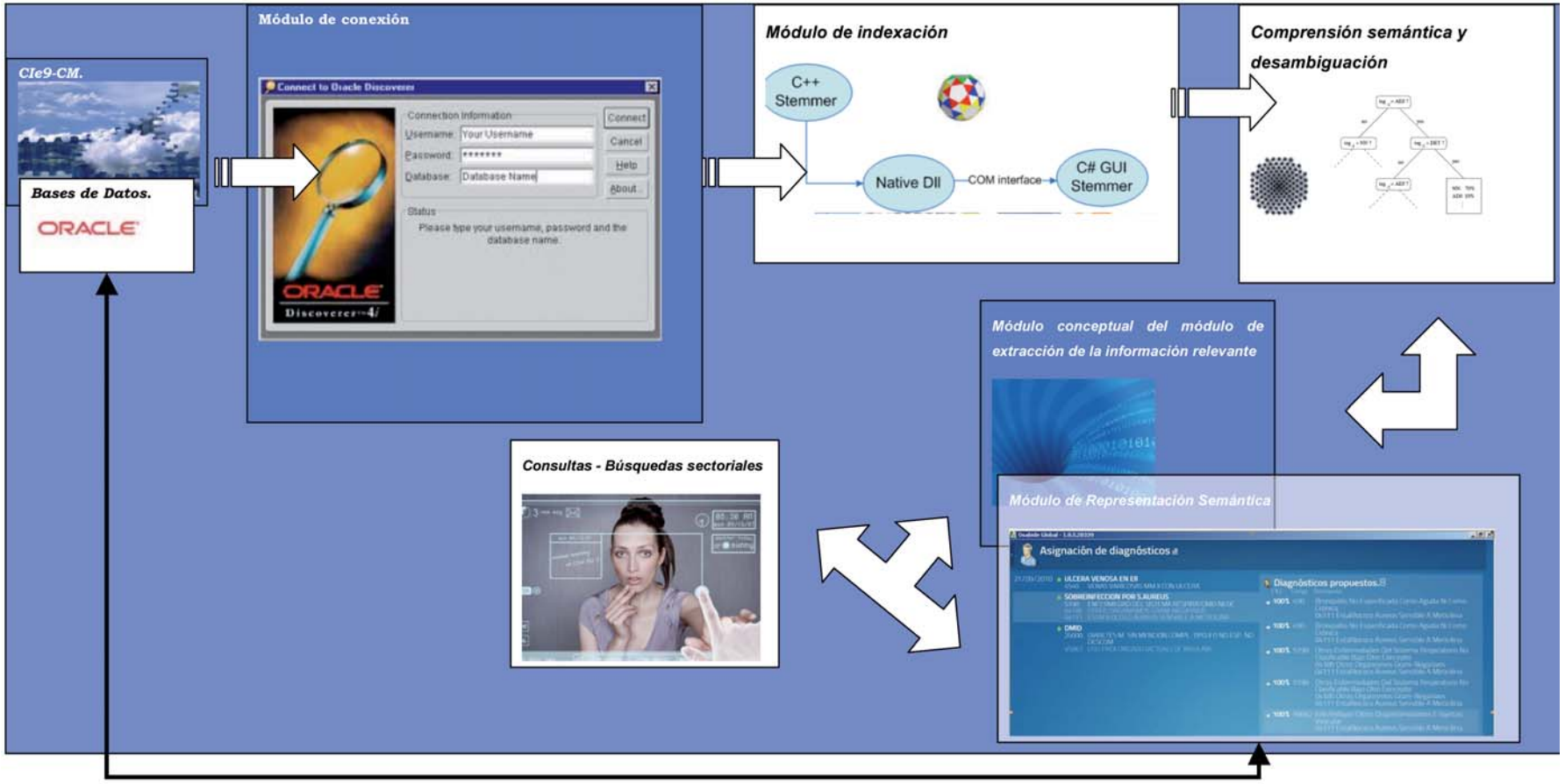
4. DISEÑO DE LA APLICACIÓN. ARQUITECTURA DE LOS MÓDULOS

En este apartado se define el software que va a soportar los módulos descritos anteriormente, después del estudio del arte realizado.

Tabla 1. **Correlación módulo-software de desarrollo**

Módulo	Software de desarrollo
Modelo conceptual del buscador semántico Framework	C#. (Contexto médico) C++. (DLL del motor de indexación) Windows Presentation Foundation (WPF)
Módulo de conexión con las fuentes externas e internas	Oracle
Módulo conceptual de indexación	Snowball
Módulo conceptual de comprensión semántica y desambiguación (tratamiento de lenguaje natural)	TreeTagger
Módulo conceptual del módulo de extracción de la información relevante	Desarrollo Propio. Según el contexto: – C#

Figura 8. **Arquitectura de módulos**



Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia
Servicio Central de Publicaciones del Gobierno Vasco

ISBN 978-84-457-3160-4