Título: Uso de la norma europea EN13606 en un sistema de historia clínica electrónica federada

Título para encabezado: Historia clínica electrónica federada

Autores:
Maldonado Segura, Jose Alberto
Robles Viejo, Montserrat
Crespo Molina, Pere
Ángulo Fernández, Carlos
Sanchis Estruch, Andrés
Saura Herranz, Alfonso

Centro:
Área de Informática Médica del Grupo BET (Bioingeniera, Electrónica y Telemedicina)
Dirección:
Montserrat Robles Viejo
ITACA
Ciudad Politécnica de la Innovación
Edificio 8-G, 3ª planta
Universidad Politécnica de Valencia
Camino de Vera S/N
Valencia 46022

Autor responsable:
Jose Alberto Maldonado Segura
Dirección: Escola Tècnica Superior d'Informàtica Aplicada
Universitat Politècnica de València
Cami de Vera S/N
Valencia 46006
Tel: 963877007-75277
E-mail: jamaldo@upvnet.upv.es

Resumen

Antecedentes. La integración de información clínica distribuida es un aspecto muy importante en el sector sanitario. El reto es desarrollar herramientas que permitan el intercambio de información entre sistemas de información sanitarios distribuidos de forma segura y conservando el significado original de los datos.

Métodos. Se utiliza la aproximación basada en mediadores y adaptadores para la integración de la información clínica en la forma de una historia clínica electrónica federada. Un mediador se puede ver como una base de datos virtual que se sitúa entre las fuentes de datos y los usuarios. Se utiliza la norma europea EN13606 del Comité Europeo de Normalización para estructurar y representar la información clínica. Por otro lado, los conceptos clínicos gestionados por el sistema se definen por medio de arquetipos. Los arquetipos son definiciones de agregados de información clínicas que tienen un significado clínico concreto y se construyen utilizando las componentes definidas en EN13606.

Resultados. Diseño e implementación en Java de un sistema historia clínica electrónica federada, denominado PANGEA. PANGEA es un middleware que permite la publicación y compartición de información sanitaria distribuida en múltiples sistemas de información en la forma de extractos de historias clínicas electrónicas compatibles con EN13606.

Conclusiones. En un proyecto de integración de información clínica la estandarización de la arquitectura de la información es básica para conservar el significado original de los datos. Por otro lado, es también importante considerar soluciones que permitan adaptarse a la continua evolución de práctica clínica. Consideramos que PANGEA cumple estos dos requisitos.

Palabras clave: historia clínica electrónica federada, integración, arquetipos, arquitectura de historia clínica electrónica

Abstract

Preliminaries. Integration is still central for health informatics. The challenge is finding how these systems can efficiently and meaningfully exchange health information about patients among several health information systems.

Methods. The mediator-wrapper approach is used to integrate health data and generate a federated electronic health record. A mediator can be seen as a virtual database, which is introduced between the data sources and the application using them. The European Standard EN13606 from the European Committee of Normalization is used for the structuring and representation of health data. On the other hand, the clinical concepts managed by the mediator are defined formally by archetypes. Archetypes are descriptions of aggregates of health data that have a clinical meaning and are defined by using eN13606 constructs.

Results. Design and implementation in Java of federated electronic health record system called PANGEA. PANGEA is a middleware that allows the publishing of health data distributed among several departmental information systems as EHCR extracts compliant with EN13606.

Conclusions. Standardization of the architecture of the clinical information is essential in order to retain the original meaning of data. Furthermore, it is also important to take into account the continuous evolution of the clinical practices, therefore the implementation of open and flexible solutions are highly recommended. We consider that PANGEA satisfies both requirements.

Palabras clave: federated electrónic health record, integration, archetypes, electronic health record architecture

1. INTRODUCCIÓN

Cada vez es más frecuente la necesidad de intercambiar información clínica entre sistemas informáticos heterogéneos. Los motivos son diversos, la atención sanitaria requiere cada vez más que profesionales sanitarios de distintas especialidades y organizaciones compartan la responsabilidad de la atención sanitaria a un mismo paciente, la movilidad de la sociedad ya sea por motivos laborales o vacacionales, los nuevos requerimientos promovidos por la sociedad de la información o para dar soporte a la investigación clínica o epidemiológica. Esto requiere obligatoriamente que los diversos sistemas puedan compartir la información clínica de los pacientes.

La telemedicina no ajena a esta necesidad de compartir información. Con la telemedicina se crean nuevos conceptos de distancia y comunicación entre médico y paciente, colaboración entre profesionales médicos o relación entre grupos de pacientes. Esta nueva medicina depende crucialmente del uso de sistemas distribuidos que hagan uso de redes de comunicación y posibiliten el intercambio de información. Estas posibilidades tecnológicas permitirán construir entornos virtuales de colaboración en medicina, que cambiarán radicalmente la manera de cooperar de los profesionales en el cuidado de la salud. Por esta razón, que uno de los aspectos tecnológicos que se tratan en la Red Temática de Telemedicina es el desarrollo de un servicio middleware (servicio común) de publicación y gestión de historias clínicas electrónicas.

La normalización de la arquitectura de la historia clínica es crucial si los datos clínicos son compartidos entre varios profesionales o instituciones. Por ello, que dentro de los trabajos de la red se ha considerado importante que el servicio de historia clínica electrónica de la plataforma global de telemedicina sea compatible con la norma europea EN13606 del Comité Técnico 251 del Comité Europeo de Normalización (CEN/TC251) [1]. Dentro de este servicio se ha considerado el desarrollo de un subservicio que permita a los repositorios de datos no normalizados servir sus datos en formatos compatibles con la norma EN113606. Es decir, que fuentes de datos clínicas ya existentes y con información relevante puedan hacer pública su información en formatos compatibles con la norma. Este artículo pretende dar una visión general del sistema informático PANGEA desarrollado por el Grupo de Informática Médica de la Universidad Politécnica de Valencia para este propósito.

2. MATERIAL Y METODOS

2.1. HISTORIAS CLINICAS ELECTRÓNICAS

En este artículo utilizaremos el término historia clínica electrónica (HCE) tal como lo define el CEN: Un registro longitudinal y potencialmente multi-institución o multinacional de la atención sanitaria de un único sujeto (paciente), creado y almacenado en uno o varios sistemas físicos con el propósito de informar en la asistencia sanitaria futura del sujeto y proporcionar un registro médico-legal de la asistencia que se le ha suministrado [2].

Pero, al tratar la integración de información clínica en soporte electrónico dispersa por múltiples sistemas, no se delimita el alcance y contenido de la historia clínica electrónica integrada. Así por ejemplo, el ámbito de la integración puede ir desde la simple compartición de datos entre dos fuentes de datos pertenecientes a una misma institución, pasando por la integración de toda la información clínica existente en un centro sanitario, hasta la integración

de toda la información de salud del paciente dispersa por todos aquellos centros donde haya recibido alguna vez atención sanitaria.

Arquitectura de información de la historia clínica electrónica

La información contenida en las historias clínicas electrónicas debe estar estructurada de alguna manera, de tal forma que se facilite su manipulación y procesamiento por un sistema informático. Esta estructura debe ser, también, adecuada tanto para el proceso de atención sanitaria como para otros posibles usos como investigación, educación, auditoría, etc. Una arquitectura de información de la historia clínica electrónica (AHCE) modela las características genéricas aplicables a cualquier anotación en una historia clínica independientemente de la organización (primaria, especializada, etc), del profesional (médico, enfermera, etc.) o especialidad. La arquitectura debe proporcionar principalmente constructores o mecanismos para capturar fielmente el significado original de la información y asegurar que la historia clínica sea comunicable. Se entiende por comunicable que la comunicación de partes de la historia clínica entre diversos profesionales, sistemas informáticos u organizaciones sea segura, y que, por tanto, se salvaguarde el significado original de los datos.

En un proyecto de integración de información clínica la estandarización de AHCE es esencial, ya que la información clínica puede ser compartida entre diversos profesionales de diversas disciplinas o puede ser transferida más allá de la organización donde fue creada. La organización destino puede ser diferente a la emisaria, puede pertenecer a otra red sanitaria o país. Sin un acuerdo en cuanto a la estructura, organización e información de contexto los datos transferidos serán difícilmente comunicables, esto es, que se salvaguarde el significado original de los datos y que por tanto los datos tengan el mismo significado tanto para el emisor como para el receptor. Ha habido diversos esfuerzos en la definición de una AHCE en Europa, cabe destacar los proyectos o grupos: Good European Health Record (GEHR) [3], el proyecto Synapses [4][5]. el Grupo de Trabajo I del Comité Técnico 251 del Comité Europeo de Normalización [CEN251] que ha dado lugar a diversas arquitecturas la última de las cuales es EN13606 [2] y el consorcio OpenEHR [6].

Norma EN13606 del CEN/TC251

Actualmente el CEN se está trabajando en su versión definitiva. EN13606 constará de cinco partes:

- Parte 1. Modelo de Referencia (*Reference Model*). Es un modelo de información genérico para la comunicación de la historia clínica de un paciente, en marzo de 2004 se hizo público el segundo borrador [2].
- Parte 2. Especificación para Intercambio de Arquetipos (*Archetype Interchange Specification*). Define un modelo de información genérico y un lenguaje para representar y comunicar instancias de arquetipos, actualmente existe un borrador público [7]
- Parte 3. Arquetipos de referencia y Lista de Términos (*Reference Archetypes and Term Lists*). Contendrá un conjunto de arquetipos que reflejen diversos requerimientos clínicos y situaciones, pretende ser una lista de partida que sirva como ejemplo para el desarrollo de nuevos arquetipos para otros dominios clínicos. Además contendrá un conjunto de listas de términos utilizadas en otras partes de la norma.

- Parte 4. Características de Seguridad (*Security Features*). Es la parte de la norma relacionada con los requerimientos de seguridad [8]
- Parte 5. Modelos de intercambio (*Exchange Models*). Un conjunto de modelos que se construyen sobre las partes anteriores y que formarían la base de un sistema de intercambio de mensajes que permitan la comunicación de toda o parte de una historia electrónica.

EN13606 esta basada en lo que se conoce como modelo dual para el diseño de la arquitectura de información para la comunicación de la HCE. En el modelo dual se distinguen dos modelos. El primero, el modelo de referencia (parte 1 de la norma), que incluye exclusivamente los conceptos no volátiles del domino que permiten describir cualquier anotación en la historia clínica. El segundo, compuesto por un conjunto de arquetipos (conformes a un modelo de arquetipos, parte 2 de la norma), que no son más que metadatos que definen por medio de restricciones sobre el modelo de referencia las características particulares de cada una de las estructuras de datos que potencialmente se necesitan para cumplir con los requisitos de información de cada grupo de profesionales, especialidad o servicio. Cuando un arquetipo restringe una componente del modelo de referencia se suele decir también que el arquetipo extiende o especializa la componente.

Los profesionales sanitarios suelen manejar un conjunto más o menos fijo de estructuras de información que representan conceptos médicos para la realización de sus actividades, por ejemplo: informe de alta, historia clínica de primaria, resultados bioquímicos, diagnóstico, etc. estos conceptos son los que se pueden formalizar por medio de arquetipos. Obviamente, la definición de arquetipos es tarea de los especialistas en el campo de interés, así por ejemplo, los patólogos puede definir arquetipos para la representación de resultados bioquímicos. Consecuencias muy interesantes del uso de un modelo dual son:

- los sistemas de información son una implementación del modelo de referencia y los conceptos del dominio, modelados como arquetipos, son definidos y usados por el sistema en tiempo de ejecución, por tanto, es más difícil que el sistema quede obsoleto.
- Permite estandarizar por separado dos procesos, la estandarización del modelo de referencia y tecnologías relacionadas, y el mucho más complejo e interminable proceso de la estandarización clínica.

Las instancias de arquetipos se ajustan a un modelo de arquetipos. Este modelo es un formalismo para expresar todos los arquetipos y está formalmente relacionado con el modelo de referencia. Esta relación se establece por medio de un conjunto de restricciones, de forma que cada concepto (clase, atributo, relación) del modelo de arquetipos se define por medio de un conjunto de restricciones sobre un concepto del modelo de referencia. Las restricciones podemos verlas como especificaciones que dictan las características que deben satisfacer las instancias de un modelo de referencia para que constituyan un concepto del domino válido. Los tipos de restricciones básicas son:

- Restricciones sobre el dominio de los atributos como declaración del valor máximo y/o valor mínimo o la enumeración de los valores aceptados.
- Restricciones sobre las relaciones agregación entre arquetipos. Esto incluye tanto la especificación de las condiciones que debe cumplir un arquetipo para que pueda estar contenido en otro como la cardinalidad de esta relación.

- Restricciones sobre la obligatoriedad de los atributos y número de ocurrencias posibles si éstos son multivaluados.

Se propone, además, un lenguaje de descripción de arquetipos (ADL) [7][9]. Tanto ADL como el modelo de arquetipos son estables pero las instancias de arquetipos individuales pueden variar a lo largo del tiempo para adaptarse a los nuevos requerimientos de la práctica clínica. El ADL contiene una sintaxis más simple denominada como dADL cuyo propósito es representar instancias de datos conformes al modelo de referencia.

Veamos con un poco más de detalle la parte primera de la norma: el modelo de referencia. EN13606-1 asume que la información contenida en la historia clínica es inherentemente jerárquica, de forma que las componentes más complejas contienen a componentes más simples. Las componentes (conceptos de negocio) definidos en EN13606-1 son:

- Carpeta (*Folder*). Representan las divisiones de más alto nivel dentro de los extractos de historia clínica, ejemplos típicos son carpeta asociada a un episodio o especialidad clínica. Permiten definir jerarquías opcionales: una carpeta puede contener a otras.
- Composición (*Composition*). Conjunto de anotaciones asociadas a una única sesión clínica o documento. Por ejemplo, informe de intervención, nota de consulta, etc. La composición es la unidad para el control de versiones de los extractos.
- Sección (*Section).* Generalmente las entradas asociadas a una única sesión clínica (composición) están agrupadas bajo diversos encabezamientos que representan las fases de la sesión, o que simplemente ayudan a mejorar la presentación y navegación por la información. Las secciones se corresponden con estos encabezamientos. Permiten definir jerarquías: una sección puede contener a otras secciones.
- Entrada (*Entry*). Representa la estructura de datos para la representación de observaciones clínicas o conjunto de observaciones (baterías de pruebas o series temporales), inferencias, acciones previstas o ya realizadas.
- Clúster *(Cluster).* La representación de una única observación o acción puede requerir una estructura compleja de datos, como una lista, tabla, o serie temporal. La clase cluster facilita los medios para representar estas agregaciones dentro de una entrada.
- Elemento (*Element).* Representa el nivel más bajo de la jerarquía de la historia clínica electrónica. Contienen un único valor que debe ser instancia de alguno de los tipos de datos definidos por el CEN [10].

Así podemos definir un arquetipo “Informe de alta” que extiende el concepto de negocio Composición. Un arquetipo puede usar otros arquetipos, de esta forma, el arquetipo “Informe de Alta” puede utilizar otros arquetipos, por ejemplo que extiendan la componente Sección del modelo de referencia, que definan conceptos como “Antecedentes”, “Recomendaciones”, “Diagnóstico”, etc.

2.2. SISTEMAS DE INFORMACIÓN FEDERADOS

Podemos entender por integración de datos el problema de combinar datos que residen en diferentes sistemas y proporcionar a los usuarios una vista unificada de éstos [11]. El desarrollo de sistemas de integración de datos surge por la necesidad de dotar a los usuarios de un acceso uniforme y transparente a varias fuentes de datos. Supongamos que no existe tal acceso, una persona que quiera acceder a datos almacenados en varios sistemas debe: saber qué bases de datos están disponibles, saber qué información hay en cada base de datos para así poder determinar su relevancia, saber descomponer la consulta en consultas parciales a

cada base de datos, conocer el modelo de cada base de datos, conocer el lenguaje de interrogación de cada base de datos y saber cómo integrar los resultados parciales para producir el resultado deseado. En conclusión, esta solución es prácticamente inviable en la mayoría de los casos.

Un buen sistema de integración debe dar a los usuarios la impresión de que están trabajando con un único sistema de información local, homogéneo y consistente. Los tres aspectos más importantes a considerar en un proyecto de integración de información son la distribución de la información, la autonomía de de las fuentes de datos y heterogeneidad [12]. El mayor obstáculo para la interoperabilidad entre varias fuentes de datos es su heterogeneidad. La principal causa es la autonomía de diseño, frecuentemente cada sistema se diseña para satisfacer los requerimientos particulares de un grupo de usuarios sin considerar requerimientos más globales. Esto se plasma en diferencias en hardware, sistemas operativos, protocolos de comunicaciones y en los sistemas de bases de datos. En relación a estas últimas podemos hablar de:

- Heterogeneidad técnica. Pueden existir diferencias en el sistema gestor de la base de datos (modelo de datos, lenguaje de interrogación, etc.), en los protocolos de acceso a datos, formatos de representación de datos o leguajes de programación.
- Heterogeneidad semántica. La heterogeneidad semántica aparece cuando existen diferencias en el significado, interpretación y uso de la información. Este tipo de heterogeneidad, es la más compleja.

En general, la heterogeneidad técnica es más fácil de resolver. Por ejemplo, el leguaje Java es independiente del hardware o del sistema operativo, lenguajes de consulta estándar, como SQL, pueden usarse para consultar las bases de datos, XML como formato común de datos y los servicios Web o CORBA se pueden utilizar como middleware para la invocación de servicios.

A lo largo de los años se han propuesto diversas arquitecturas para la integración de datos, una de las más empleada son los sistemas de información federados. Podemos entender por sistema de información federado aquel que está compuesto por un conjunto fuentes de datos, que llamaremos fuentes componentes, que son heterogéneas, distribuidas y autónomas pero que ceden parte de su autonomía para dar acceso no sólo a los usuarios locales de cada base de datos sino a un conjunto de usuarios globales que se conectan a través de alguna clase de red de comunicación. En un sistema federado las fuentes de datos componentes siguen conservando gran parte de su autonomía pero cooperan para satisfacer los requerimientos de información de los usuarios globales de la federación. Se denominan sistemas de información federados fuertemente acoplados a aquellos que facilitan un esquema federado (también conocido como esquema global). Este esquema contiene una vista integrada y reconciliada de los esquemas de las fuentes de datos integradas y se expresa en un modelo de datos común o canónico (MDC). Cada modelo de datos local se mapea al MDC, es decir, los datos que contienen se convierten a estructuras de datos conformes con el MDC. Dentro de este tipo se encuadran los mediadores.

El término mediador (mediator) fue introducido por Wiederhold en [13] y desde entonces se utiliza en la literatura sobre técnicas y proyectos de integración de datos. Los mediadores son programas informáticos especializados que obtienen la información a partir de una o más fuentes de datos o de otros mediadores y proporcionan información a los componentes que están por encima (otros mediadores) y a los usuarios externos del sistema [14]. En un

mediador las fuentes de datos están “envueltas” por una capa de software, denominada adaptador o *wrapper*, el cual traduce entre el lenguaje, modelos y conceptos de la fuente de datos y el lenguaje, modelo y conceptos utilizados en el mediador. Un mediador ofrece una vista unificada e integrada de la información que se encuentra almacenada en las diversas fuentes de datos. El esquema global (federado) proporciona una vista virtual, integrada y reconciliada de las fuentes de datos subyacentes. El esquema global se suele construir de acuerdo con las necesidades de información de los usuarios globales. Por ello, que se puede entender a los mediadores como servicios que se construyen y se ponen a disposición de los clientes. Generalmente los mediadores solo permiten la lectura de datos y no la escritura, es decir, no es posible añadir o modificar datos a través de la vista integrada, sino que para ello se debe utilizar los sistemas componentes de la federación.

3. RESULTADOS: El sistema de integración PANGEA

3.1. GENERALIDADES

Básicamente PANGEA es un mediador que permite construir una historia clínica electrónica federada (HCEF), término acuñado por el proyecto Sypnases [4]. El adjetivo *federada* no es gratuito sino que se refiere al uso de las ideas de los sistemas de información federados para el desarrollo de sistemas de historias clínicas electrónicas. Una HCEF es una historia clínica electrónica virtual, por virtual se entiende que no se encuentra almacenada en un sistema de bases de datos sino que se construye al vuelo y bajo demanda a partir de información distribuida en varios sistemas informáticos, posiblemente heterogéneos entre si, pertenecientes a una o varias organizaciones. Una HCEF puede englobar más o menos información clínica sobre un paciente, desde la simple integración de algunos repositorios de datos clínicos hasta englobar toda la información disponible independientemente de la institución donde se encuentre. Cada fuente de datos puede permitir el acceso a toda o parte de la información que contiene, son los responsables de los datos los que controlan qué información se comparte con el resto de sistemas que forman la federación. Para el desarrollo de un sistema de HCEF es crucial el uso de un modelo de datos consensuado para la representación de la información clínica, es decir, una arquitectura de información para la comunicación de la historia clínica. Este modelo debe salvaguardar el significado original de los datos, de forma que el destinatario pueda entender correctamente la información enviada.

3.2. MODELO DE DATOS

En un mediador existe un esquema global (federado) que contiene una vista unificada de la información repartida por las fuentes de datos. En un sistema de historia clínica electrónica federada este esquema global describe la información clínica manejada y publicada por el sistema. El esquema federado debe expresarse en un modelo de datos canónico (MDC) para ocultar los formatos particulares de las fuentes de datos y por tanto facilitar una representación única de los datos. Consideramos que el MDC a utilizar para la descripción de la información clínica debe poseer las siguientes características:

- Completitud. En la metodología dual los extractos de historias clínicas son instancias del modelo de referencia. Por tanto, el modelo de datos debe tener la suficiente capacidad expresiva para poder representar cualquier instancia de los conceptos (clases) definidos en el modelo de referencia. Como el modelo de referencia, se describe por medio de un diagrama de clases, por ejemplo en UML, el modelo de

datos debe poseer constructores que permitan representar modelos de datos orientados a objetos.

- Simplicidad. El MDC debe ser lo suficientemente simple para ser fácilmente entendido y utilizado por los diseñadores de arquetipos, esto implica que exista una representación gráfica intuitiva de las estructuras de datos.
- Compatibilidad. Es importante que el modelo facilite la representación de los datos utilizando las sintaxis candidatas para la comunicación de los extractos de historias clínicas como XML o dADL. Es decir, que el modelo de datos sea compatible con los modelos de datos subyacentes a ambas sintaxis.

El modelo de datos utilizado para la representación de extractos de historias clínicas electrónicas compatibles con la norma europea EN13606 se basa en árboles etiquetados con referencias donde los datos están asociados a los nodos. El modelo se adapta perfectamente a EN13606, recordemos que la norma asume una organización jerárquica, es decir en forma de árbol, de los extractos de historias clínicas electrónicas. El modelo de datos utilizado es similar a otros modelos de datos utilizados para modelar datos semiestructurado o XML (eXtensible Markup Language o Lenguaje Extensible de Marcado) como [15][16][17]. A su vez es compatible con XML. Cabe destacar, que en el fondo, XML no es más que una convención para la representación de árboles etiquetados como texto, esto es, como una secuencia de caracteres y que en si mismo no es un modelo de datos.

Utilizamos los nodos internos del árbol para representar atributos tanto simples como complejos y cuyas etiquetas provienen del modelo de referencia, mientras que las hojas contienen los datos y, por tanto, están etiquetadas con valores pertenecientes a los tipos básicos. Las referencias son necesarias, por ejemplo, para modelar las relaciones de contención por referencia presentes en el modelo de referencia de EN13606, como la existente entre los objetos de tipo carpeta (FOLDER) y los de tipo composición (COMPOSITION) donde las carpetas no contienen físicamente a los objetos de tipo composición sino que almacenan sus identificadores. El modelo permite que los nodos estén tanto ordenados como no. El orden es necesario para poder representar estructuras como listas o documentos donde el orden los elementos es relevante.

3.3. ARQUETIPOS

Los arquetipos son la base de la integración en PANGEA. Su propósito es hacer públicos los datos contenidos en los sistemas de información a integrar y al mismo tiempo ocultar su heterogeneidad, es decir, forman un nivel semántico sobre las bases de datos y sirven para asociar a los datos almacenados en éstas una semántica clínica específica. Por tanto, en nuestro sistema la verdadera integración se realiza a nivel de metainformación en vez de a nivel de datos. A su vez los usuarios extraen información por medio de la instanciación de uno o varios arquetipos. Resumiendo los arquetipos definen los conceptos clínicos que se comparten y los extractos de historias clínicas que se facilitan a las aplicaciones clientes son siempre instancias de arquetipos.

En nuestro caso, donde se pretende utilizar un modelo dual de arquitectura para la comunicación de la historia clínica electrónica es necesario adaptar esta metodología para su uso en un sistema de historia clínica electrónica federada. Las principales consecuencias de esto son:

- Los arquetipos deben ser las entidades constitutivas del esquema global, es decir, los conceptos que se representan en el esquema federado deben ser arquetipos tal como éstos se entienden en la norma europea.
- El esquema federado no se construye por un proceso de integración de esquemas en tiempo de diseño sino que su construcción es dinámica. El esquema global crece a lo largo del tiempo, por ejemplo tras la incorporación de una nueva fuente de datos a la federación o por nuevos requerimientos de información de los usuarios.
- La interrogación de datos debe estar basada en arquetipos.

Para PANGEA hemos desarrollado un sistema de tipos (definición de esquemas) para el modelo de datos anteriormente descrito. Los arquetipos se modelan como un tipo sobre el modelo de datos, por tanto, cada arquetipo define un conjunto de extractos de HCE que comparten un conjunto de características estructurales y de etiquetado. Este sistema de tipos permite también la construcción del esquema federado, en efecto, el esquema federado no es más que un conjunto de tipos. Los tipos se basan en el uso de predicados que definen el dominio de los valores atómicos y en expresiones regulares para la especificación de la estructura permitida de los extractos [15]. La Figura 1 describe gráficamente la relación existente entre los arquetipos y los tipos sobre los árboles de datos.

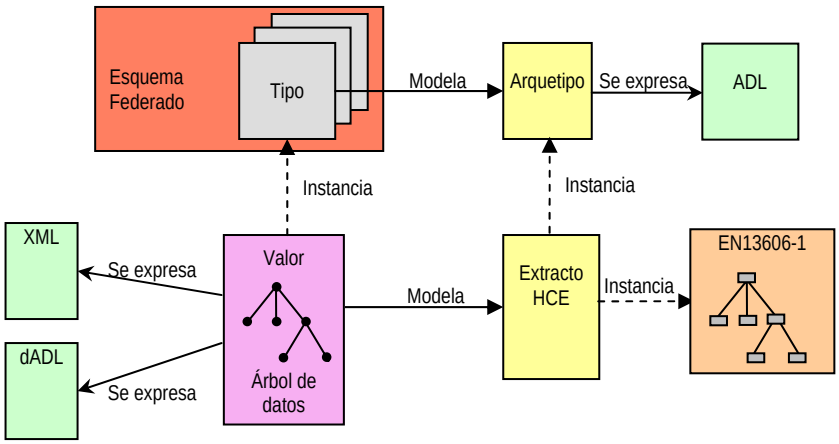

Figura 1. Relación entre tipos y arquetipos

Una duda puede aparecer, ¿por qué no utilizar los mecanismos para la especificación de esquemas (tipos) para XML de la W3C como XML Schema para la definición de arquetipos?. Recordemos que XML Schema es un metalenguaje desarrollado por la W3C para especificar clases de documentos XML y, por tanto, permite especificar qué documentos son válidos para una aplicación en particular. Su uso está muy extendido y existen multitud de herramientas asociadas. Por tanto, su uso como mecanismo para la representación de arquetipos parece recomendable. El problema estriba en su falta de potencia, por ejemplo, en XML-Schema los tipos que aparecen en la expresión regular asociada a un tipo no pueden tener asociados la misma etiqueta, lo cual tiene graves consecuencias para la descripción de arquetipos con XML-Schema. La principal es la limitación en cuanto a la composición de arquetipos: un arquetipo puede tener como máximo un “subarquetipo” de cada tipo de componente definido en el modelo de referencia. Así por ejemplo, un arquetipo “carpeta” únicamente puede tener como máximo un arquetipo “composición”. Por tanto, no es posible describir la mayoría de los arquetipos, incluso los más sencillos. En [18] se puede encontrar un estudio detallado del uso de diversos sistemas de tipos de XML para la descripción de arquetipos.

3.4. ENLACE DE LOS ARQUETIPOS CON LOS DATOS

Ya que la información clínica reside en las fuentes de datos federadas, se debe definir algún tipo de correspondencia entre los arquetipos y los esquemas de éstas. En la teoría de las bases de datos, las vistas definen un subconjunto de la información contenida en una base de datos. Por tanto, un arquetipo se puede considerar una vista que facilita la abstracción e integración de los datos clínicos. Además, las vistas (arquetipos) deben presentar los datos como extractos de HCEs compatibles con EN13606 en forma de documentos XML. Para conseguir esto y debido a la heterogeneidad entre el modelo relacional (predominante en el sector sanitario) y XML, las vistas deben proporcionar además de un conjunto de consultas que extraigan la información relevante, información de mapeo que permita estructurar y etiquetar el documento XML resultante. La publicación de datos relaciones como XML no es una tarea trivial; los datos relaciones no están anidados, están normalizaron en varias tablas y el esquema es a menudo propietario. Por el contrario, XML es anidado, no está normalizado y su esquema es público (en nuestro caso EN13606-1). La publicación en forma de documentos XML implica, combinar las tablas, seleccionar los datos que aparecerán en el documento XML, enlazar los atributos de las tablas con los elementos o atributos del XML y por último estructurar el documento XML resultante.

En PANGEA las correspondencias entre los arquetipos y los esquemas de las bases de datos se establecen haciendo corresponder a un atributo de un arquetipo una expresión que englobe uno o varios atributos de las fuentes de datos o constantes. A estas correspondencias las denominamos correspondencias de valores. A continuación se muestran dos ejemplos de estas correspondencias:

Altura.Resultado_de_valor_numérico.valor_numérico← Consultas_externas.medidas.altura*100
Altura.Resultado_de_valor_numérico.Unidad_de_medida←”cm”

La parte izquierda de la correspondencia especifica un camino dentro de la definición de arquetipo donde los nombres de los arquetipos se muestran en cursiva. En la primera correspondencia al atributo “valor_numérico” del atributo complejo “resultado_de_valor_numérico” del arquetipo *Altura* se le asocia el valor contenido en el atributo “Altura” de la tabla “Medidas” de la base de datos “Consultas_externas” multiplicado por 100. La segunda correspondencia simplemente asigna la constante “cm” al atributo “Unidad_de_medida”. Obviamente, el arquetipo Altura debe restringir algún concepto de negocio definido en el modelo de referencia que tenga como atributo a “Resultado_de_valor_numérico”.

A partir del conjunto de correspondencias de valores, la estructura inducida por los arquetipos y el modelo de referencia, PANGEA es capaz de generar un conjunto de consultas candidatas que permiten extraer la información relevante para la instanciación del arquetipo para un paciente en particular. A la hora de generar una consulta candidata se ha tenido en cuenta los siguientes aspectos:

- Debemos mantener todas las relaciones existentes entre los datos, típicamente expresadas en forma de claves ajenas, en las fuentes de datos relacionales.
- No debemos perder información.

Para este fin se ha utilizado el concepto de disyunción completa de un conjunto de relaciones (tablas) en el caso del modelo de datos relacional. La disyunción completa se define como la

máxima información sin redundancias y manteniendo todas la relaciones existentes entre los datos que puede ser obtenida a partir de un conjunto de relaciones. Como se demuestra en [19] la disyunción completa de un conjunto de relaciones es única y en la mayoría de las ocasiones se pueden calcular por medio de expresión que solo contiene concatenaciones externas completas (full outer join). La expresión obtenida se puede simplificar si tenemos en cuenta las restricciones impuestas por los arquetipos sobre los datos, las consultas de los usuarios, las propiedades de las claves ajenas y que no toda la información contenida es relevante: en efecto, solo estamos interesados en aquella información para la cual podamos determinar a qué paciente pertenece. Generalmente, las consultas resultantes solo hacen uso de concatenaciones internas y concatenaciones externas a izquierdas (left joins), las cuales pueden ser ejecutadas por la mayoría de los sistemas de bases de datos.

Esta aproximación facilita la labor de definir la información de mapeo entre los arquetipos y las fuentes de datos relacionales ya que es más fácil para el diseñador de arquetipos definir qué campo de la base de datos o expresión se debe utilizar para poblar un atributo del arquetipo que especificar el conjunto de consultas, posiblemente complejas, necesarias para extraer la información. Las consultas candidatas generadas pueden y deben ser revisadas por el diseñador de arquetipos para comprobar su validez y ser modificarlas en caso de resultar erróneas. Las correspondencias de valores tienen implícitamente información estructural que facilitan en gran medida la conversión entre los datos relaciones y el documento XML. Así es posible derivar automáticamente el conjunto de atributos de las relaciones que unívocamente identificar una instancia de un arquetipo o de un atributo complejo y por tanto estructural el documento XML resultante [20]. Como consecuencia la conversión relacional-XML es sencilla y puede realizarse utilizando una herramienta propia (este es el caso de PANGEA) o generar automáticamente algún tipo de especificación (script) para alguna herramienta comercial o de libre distribución.

3.5. SISTEMA INFORMÁTICO

El sistema informático desarrollado enteramente en Java, véase figura 2, se sitúa entre los usuarios y las bases de datos a integrar. Por tanto, se puede considerar como una forma de middleware, que recupera, bajo demanda de los usuarios, la información relevante sobre el paciente y la entrega a los usuarios. El sistema genera un documento XML que contiene la información clínica demandada y que está estructurada según el EN13606-1. El servidor de HCE presenta una interface sencilla que se hace pública mediante servicios web (Web Servicies). Esto dota de gran independencia a la hora de desarrollar aplicaciones clientes (por ejemplo estaciones clínicas) que serán las que presenten los resultados a los usuarios finales.

Otra de las componentes principales del sistema es el servidor de metainformación, el cual gestiona el diccionario de datos donde se almacena toda la información necesaria para el funcionamiento del sistema. El diccionario de datos contiene información sobre usuarios y sus permisos, las bases de datos conectadas al servidor, los esquemas de las bases de datos, qué objetos de las bases de datos conectadas pueden ser accedidos por las aplicaciones cliente, la definición y versiones anteriores de los arquetipos, los enlaces de éstas con los esquemas de bases de datos, las relaciones entre arquetipos y sinónimos.

El servidor de historias clínicas federadas es la interfaz el sistema con los usuarios, por tanto, gestiona el esquema federado (conjunto de arquetipos). Utiliza otros dos servidores para realizar su tarea, el servidor de control de acceso evita que usuarios no autorizados puedan acceder a los datos y el servidor de identificadores de pacientes que permite identificar

pacientes idénticos entre diferentes sistemas cuando no existe un identificador de paciente universal. Se han desarrollado también, tres aplicaciones visuales para la gestión de la metainformación del sistema: gestor de permisos, el editor de arquetipos y el gestor de esquemas de bases de datos. La primera tiene como propósito gestionar los permisos de los usuarios en relación con el uso, instanciación, creación y edición de arquetipos. La segunda es una herramienta que facilita la creación de nuevos arquetipos ya sea empezando desde cero o reutilizando ya existentes. Permite, además, modificar arquetipos ya existentes conservando las versiones previas y establecer las correspondencias entre los arquetipos y las fuentes de datos y validar las consultas generadas. La última es una herramienta para visualizar, gestionar y completar los esquemas de las bases de datos conectadas al servidor.

4. CONCLUSIONES Y TRABAJOS FUTUROS

La necesidad de acceder a información clínica de manera uniforme repartida en múltiples fuentes de datos heterogéneas no es ajena a la telemedicina. Por ello, alguno de los aspectos tecnológicos que se tratan en la Red Temática de Telemedicina es el desarrollo de un servicio middleware (servicio común) de publicación y gestión de historias clínicas electrónicas basado en normas europeas de informática médica. Dentro de este servicio se ha considerado el desarrollo de un sistema informático que permita a los repositorios de datos no normalizados puedan servir sus datos en formatos compatibles con la norma EN13606.

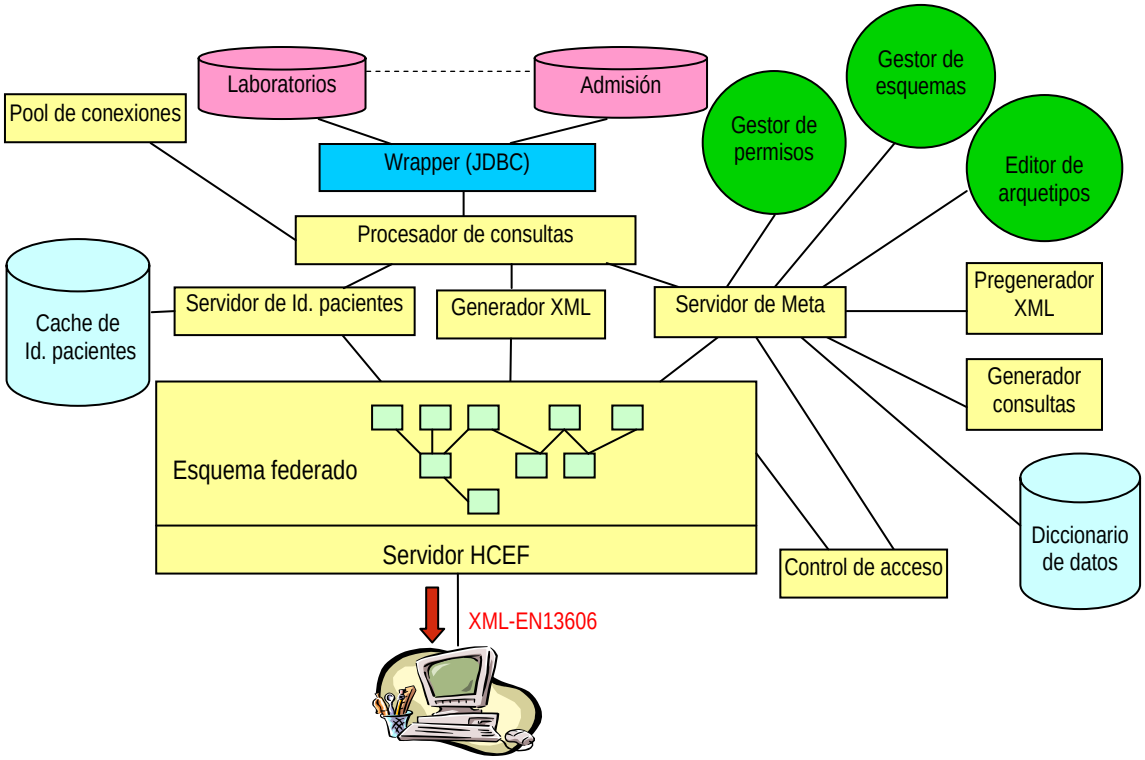

Figura 2. Arquitectura de PANGEA.

Desde el punto de vista técnico PANGEA es un mediador (middleware) que permite construir una historia clínica electrónica federada basada en la norma europea EN13606 a partir de la información clínica dispersa en múltiples repositorios de datos heterogéneos, ya existentes y no basados en la norma. La historia clínica electrónica federada ofrecida se basa en un modelo dual de arquitectura de historia clínica electrónica y por tanto los arquetipos son las entidades que componen el esquema federado ofrecido por PANGEA. Los arquetipos que son la base

del sistema de integración ya que definen los conceptos clínicos manejados por el sistema y a la vez ocultan la heterogeneidad de las fuentes de datos, es decir, forman un nivel semántico sobre las bases de datos y sirven para asociar a los datos almacenados en éstas una semántica clínica específica. El mapeo entre los arquetipos y las fuentes de datos se realiza siguiendo una política de *especificar y generar*, se especifica un conjunto de correspondencias de alto nivel entre ambos y se genera un mapeo candidato que debe ser validado por el diseñador del arquetipo. Los usuarios solicitan información instanciando uno o varios arquetipos para un paciente en particular. El sistema se encarga de obtener la definición e información de mapeo con las fuentes de datos del arquetipo para continuación construir un documento XML conforme a EN13606-1 que contiene el extracto de la historia clínica electrónica solicitada.

Actualmente se está trabajando en la incorporación de fuentes de datos no relacionales tales como imágenes, sistemas de ficheros o sistemas de mensajería, la implementación completa de EN13606-1 y en el aumento de la potencia del procesador de consultas para que permita realizar consultas que den soporte a la investigación clínica o epidemiológica. Hemos abierto una nueva línea de trabajo centrada en dotar de mayor semántica a los datos, en concreto la incorporación de un servidor de terminología que ayude en la definición de arquetipos y en descripción y la gestión de la información clínica basadas en tecnologías de la web semántica.

5. BIBLIOGRAFIA

1. www.centc251.org
2. CEN/TC 251. Health Informatics-Electronic Healthcare Record Communication- Part 1: Reference Model. prEN 13606-1 2nd working draft, documento N04-012, 2004. Disponible en: http://www.centc251.org/TCMeet/doclist/TCdoc04/N04-012prEN13606-1_2WD.doc [visitado el 13 de Enero de 2005].
3. http://www.chime.ucl.ac.uk/work-areas/ehrs/GEHR
4. Grimson, W., Berry, D., Grimson, J., Stephens, G., Felton, E., Given, P., O'Moore, R. Federated healthcare record server-the Synapses paradigm. International Journal of Medical Informatics 1998; 52:3-27.
5. Grimson, J., Grimson, W., Berry, D., Stephens, G., Felton, E., Kalra, D., Toussaint, P., Weier, O. W. A CORBA-based integration of distributed electronic healthcare records using the synapses approach. IEEE Transactions on Information Technology in Biomedicine 1998; 2(3)-124-138.
6. http://www.openehr.org
7. CEN/TC 251. Health Informatics-Electronic Healthcare Record Communication Part 2: Archetype Interchange Specification. prEN 13606-2, 2003. Disponible en: http://www.centc251.org/WGI/N-documents/N03-20prEN13606-2_(E)_v0.2.doc [visitado el 13 de Enero de 2005].
8. CEN/TC 251. Health informatics-Electronic health record communication Part 4: Security requirements and distribution rules, 2003. Disponible en: http://www.centc251.org/WGI/N-documents/N03-21%20EN_13606-4__E_-v04.pdf [visitado el 13 de Enero de 2005]
9. OpenEHR (2004). Archetype Definition Language (ADL), Revision 1.2. Disponible en: http://www.openehr.org/drafts/ADL-1_2_draftF.pdf [visitado el 13 de Enero de 2005]
10. CEN/TC 251. Health Informatics-Data Types, CEN/TS 14796. Disponible en: [http://www.centc251.org/TCMeet/doclist/TCdoc03/N03-042prCEN-TS-14796RevisedFinalDraft.pdf [visitado el 13 de Enero de 2005]

11. Lenzerini, M. Data Integration: A Theoretical Perspective. Proceedings of the 21st ACM SIGACT-SIGMOND-SIGART Symposium on Principles of Database Systems 2202; 233-246.
12. Sheth, A.P., Larson, J.A. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases, ACM Computing Surveys 1990; 22(3)-183-236.
13. Wiederhold, G. Mediators in the Architecture of Future Information Systems. Computer 1992; 25(3)-38-49.
14. Ullman, J. D. Information integration using logical views. Proceedings of the 6th International Conference on Database Theory (ICDT'97) 1997; 19-40.
15. Kuper, G. M., Simeon, J. Subsumption for XML types. Proceedings of the 8th International Conference on Database Theory (ICDT'01) 2001: pp. 331-345.
16. Abiteboul, S., Cluet, S., Milo, T. Correspondence and translation for heterogeneous data. Theoretical Computer Science 2002; 275(1-2) 179-213.
17. Milo, T., Suciu, D. Type inference for queries on semistructured data. Proceedings of the 18th ACM SIGACT-SIGMOND-SIGART Symposium on Principles of Database Systems 1999: 216-226.
18. Tun, Z., Bird, L. J., Goodchild, A. (2002). Validating electronic health records using archetypes and XML. Technical Report of the TITANIUM Project, disponible en: http://titanium.dstc.edu.au/papers/acsc2002.pdf [visitado el 14 de Enero de 2005].
19. Rajaraman, A., Ullman, J. D. Integrating information by outerjoins and full disjunctions. Proceedings of the 15th ACM SIGACT-SIGMOND-SIGART Symposium on Principles of Database Systems 1996; 238-248.
20. Shanmugasundaram, J., Shekita, E., Barr, R., Carey, M., Lindsay, B., Pirahesh, H., Reinwald, B. Efficiently publishing relational data as XML documents. The VLDB Journal 2001; 10(2-3)133-154.

AGRADECIMIENTOS

El trabajo presentado está parcialmente financiado por el Ministerio de Sanidad a través de la Red Temática de Investigación Cooperativa en nuevos servicios de salud basados en telemedicina y por el Ministerio de Educación y Ciencia proyecto TSI2004-06475-C02-01 dentro del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica.