# NORMALIZACIÓN DE UMLS PARA USO EN APLICACIONES MÉDICAS

*R. Lozano; S. Saiz; F. Geva; X. Pastor*
*Informática Médica. Corporación Sanitaria Clínic. Barcelona*

## RESUMEN

El proyecto SCOPE pretende hacer accesible la consulta on-line de una revista de gastroenterología (G&H) en 4 idiomas, manteniendo una alta especificidad respecto a la información recuperada, para que sólo se entregue al usuario información relativa al tema de interés manifestado por éste, eliminando toda información no relevante.

Los aspectos relacionados con la terminología han sido abordados mediante la utilización del Unified Medical Languaje System (UMLS), un proyecto de la National Library of Medicine (NLM) que agrupa actualmente a más de 60 clasificaciones internacionales sobre el dominio médico, como marco terminológico de referencia.

No obstante, la utilización de UMLS no está exenta de inconvenientes. El sistema es entregado como un conjunto de enormes ficheros ASCII, no normalizados, con gran cantidad de datos redundantes de muy difícil utilización y sin herramientas para la búsqueda de conceptos e información relacionada. Para solventar este escollo se ha procedido al diseño y construcción de una base de datos relacional normalizada para UMLS y a su carga a partir de los ficheros ASCII. Disponemos así de una herramienta muy eficiente a la hora de encontrar conceptos médicos relacionados con una palabra o conjunto de palabras, independientemente del idioma utilizado.

SCOPE (Structuring Content for On-line Publishing Environments) está financiado por la Unión Europea en el programa e-Content [EDC-22016Y1C1DMAL2]

## INTRODUCCIÓN

El proyecto SCOPE (Structuring Content for On-line Publishing Environments)(1) pretende fomentar la explotación de contenidos científicos en un entorno multilingüe, para lo que contempla hacer accesible la consulta on-line de una revista de gastroenterología (G&H) en 4 idiomas. Un requerimiento del proyecto es la alta especificidad respecto a la información recuperada, para que sólo se entregue al usuario información relativa al tema de interés manifestado por éste, eliminando toda información no relevante.

Un aspecto crucial es la representación de la información de modo que los contenidos de la revista sean interpretables por un ordenador, soportando varios idiomas. Ello implica primero el diseño y construcción de un sistema capaz de albergar una representación de los contenidos susceptible de ser interrogado, que se ha plasmado en la utilización de RDF(2;3) para la definición de una ontología que represente los contenidos. Luego es preciso identificar los conceptos a utilizar y los posibles modos de referirse a ellos. La solución adoptada para este segundo requerimiento ha sido la de utilizar Unified Medical Languaje System (UMLS)(4), un proyecto de la National Library of Medicine (NLM) que recoge en la actualidad a más de 700.000 conceptos diferentes de más de 60 clasificaciones internacionales sobre el dominio médico, como marco terminológico de referencia.

UMLS está organizado a través de tres componentes:

1. El Metathesaurus: Recoge los conceptos que aparecen en alguno de los vocabularios que integra. En el Metathesaurus se mantienen los significados, atributos, conexiones jerárquicas y otras relaciones entre términos presentes en sus vocabularios de origen, y se añade, por otra parte, nuevas relaciones entre conceptos y términos de distintos vocabularios de origen.

2. El Lexicón: Recoge información léxica necesaria para el procesamiento de lenguaje natural.

3. La Semantic Network: Su propósito es proveer de una categorización consistente de todos los conceptos representados en el Metathesaurus, así como de relaciones entre los mismos (enlaces is-a).

Para el proyecto SCOPE hemos utilizado solamente el Metathesaurus, en un doble sentido. Primero, como fuente de conceptos entre los que escoger aquellos que se vaya a necesitar para construir la ontología. Segundo, para resolver el problema de la multilingualidad. UMLS recoge descripciones en múltiples lenguas, lo que permite crear la ontología independientemente de cualquier idioma y utilizar UMLS como una especie de "traductor".

UMLS es considerado cada vez más como un estándar en terminología médica, permitiendo la integración tanto de aquellas aplicaciones basadas en UMLS como de las que utilizan sistemas de clasificación considerados dentro de UMLS, por lo que cada vez es más utilizado(5-8).
La búsqueda de conceptos en UMLS exige un sistema ágil de exploración entre el enorme número de conceptos y descripciones relacionadas que contempla, así como contar con una mínima garantía respecto a la consistencia e integridad de la información. Sin embargo la NLM entrega UMLS como un conjunto de ficheros ASCII de enorme tamaño (algunos ocupan más de 1 Gb), con una estructura realmente compleja y sin herramientas para su explotación, lo que hace muy difícil su utilización directa. Como veremos además, tampoco la consistencia e integridad de la información está garantizada. Se impone por tanto la construcción de un sistema que aporte la funcionalidad requerida. Aunque existen algunos trabajos relativos a la semántica de UMLS(9;10), no ocurre lo mismo respecto al contenido del Metathesaurus.

## MATERIAL Y MÉTODOS

El objetivo básico es el de diseñar y construir una base de datos relacional en 3ª forma normal que permita almacenar el contenido del Metathesaurus de UMLS así como una eficiente exploración del mismo.

El Metathesaurus de UMLS está organizado sobre la base de tres niveles:

- Concepto: Recoge la conceptualización en sí. Por ejemplo el concepto de fibrilación auricular.
- Término: Hace referencia a los distintos modos básicos de referirse a un concepto. Por ejemplo, son términos distintos del concepto de fibrilación auricular: fibrilación auricular y fibrilación atrial.
- String: Variaciones sobre los términos, por ejemplo el término en singular y en plural: fibrilación auricular y fibrilaciones auriculares son diversas strings relacionadas con el término fibrilación auricular.

Además, el Metathesaurus de UMLS recoge un gran número de relaciones entre los conceptos, por ejemplo todas las relaciones jerárquicas que se dan en las clasificaciones de origen, como la ICD 9 CM.

Aunque la documentación aportada por la NLM se refiere a los ficheros que entrega como "formato relacional ASCII", la realidad es que, aunque efectivamente se pueden considerar relaciones, no se hallan normalizados y presentan numerosas redundancias. La metodología seguida ha consistido en una aproximación clásica en tres niveles:

1. Diseño de un modelo conceptual utilizando el modelo Entidad/Interrelación.
2. Especificación de un modelo lógico relacional sobre la base del modelo conceptual.
3. Implementación del modelo físico sobre un gestor de bases de datos relacional, en este caso sobre el Adaptive Server Enterprise de Sybase? .

Como último paso figura la carga del sistema a partir de los ficheros originales.

Realmente, y como cabía esperar, el proceso ha resultado bastante cíclico e iterativo. Sobre un modelo e implementación determinados ha sido preciso intentar cargar los datos para descubrir determinadas relaciones o características hasta entonces ocultas, que han supuesto la modificación del modelo conceptual y las modificaciones en cascada pertinentes. No obstante haremos una exposición secuencial de la metodología empleada.

El proceso de diseño del modelo conceptual ha resultado como sigue:

1. Identificación de las entidades participantes y sus atributos: los ficheros originales presentan un alto grado de mezcla de lo que se pueden considerar entidades conceptuales distintas y la documentación no aporta excesiva luz al respecto. Por lo tanto, el primer paso ha consistido en determinar qué grupos de atributos podían identificarse como relacionados entre sí según algún tipo de dependencia. A falta de una documentación más explícita ha sido necesario recurrir muchas veces a examinar el contenido de los ficheros.
2. Reconocer las claves primarias de las entidades. Mientras que en el caso de entidades dotadas de un código en el sistema original (como el código de concepto respecto a la entidad concepto) esta tarea es trivial, en otras entidades ha resultado sumamente difícil establecer qué conjunto de atributos identificaban unívocamente a una ocurrencia o cual de las claves candidatas era la más adecuada.
3. Descubrir las relaciones entre las entidades, describiendo el grado y cardinalidad de las mismas, dentro del contexto del sistema, añadiendo las claves foráneas precisas.

Aunque los modelos Entidad/Interrelación producen modelos relacionales prácticamente normalizados de forma automática, cada propuesta ha sido cuidadosamente analizada para garantizar el cumplimiento de la 3º forma normal. Su finalidad es reducir las inconsistencias y redundancias de los datos, facilitar el mantenimiento y evitar las anomalías en las manipulaciones de datos.

El procedimiento de normalización consiste en someter a las tablas que representan las entidades del modelo conceptual a un análisis formal para ver si cumplen, o no, las restricciones necesarias que aseguren evitar los problemas citados con anterioridad. A mayor nivel de normalización, mayor calidad en la organización de los datos y menor peligro para la integridad de los mismos.

A modo de ejemplo podemos ver el proceso seguido con uno de los ficheros básicos originales, MRCON, que presenta una única entrada por cada significado de cada string única en el Metathesaurus. La documentación aporta la siguiente información:

***Fichero origen : Concept Names MRCON ( CUI, LAT, TS, LUI, STT, SUI, STR, LRL )***

**CUI Unique identifier for concept**
**LAT Language of Term**
**TS Term status**
**LUI Unique identifier for term**
**STT String type**
**SUI Unique identifier for string**
**STR String**
**LRL Least Restriction Level**
**Sample Records**

C0002871|ENG|P|L0002871|PF|S0013742|Anemia|0|
C0002871|ENG|P|L0002871|VP|S0013787|Anemias|0|
C0002871|ENG|P|L0002871|VC|S0352787|ANEMIA|0|
C0002871|ENG|P|L0002871|VC|S0414880|anemia|0|
C0002871|ENG|P|L0002871|VO|S0470197|Anemia,NOS|3|
C0002871|ENG|S|L0280031|PF|S0803242|Anaemia|3|

Como puede apreciarse, hay gran cantidad de información redundante, apareciendo repetido el código del concepto (CUI) por cada entrada, cuando es un dato dependiente del término.

Tras el proceso de análisis y diseño ha quedado del siguiente modo:

1. **concept (cui, validity)**
   **Recoge los conceptos existentes en el Metathesaurus.**
   - cui: identificador único de concepto. Constituye la clave primaria.
   - validity: Fecha de validez de un concepto. Recoge el año hasta el cual el concepto es válido.
2. **term (lui, cui, ts)**
   **Recoge los términos existentes.**
   - lui: identificador único del término. Constituye la clave primaria.
   - cui: identificador del concepto al que hace referencia el término.
   - ts: estado del término. Indica si se trata del término preferido para referirse al concepto, si es un sinónimo, etc.
3. **string (sui, lui,str,lat,lrl)**
   Recoge las cadenas de caracteres (strings)
   - sui: Identificador único por string. Constituye la clave primaria.
   - lui: Identificador del término al que hace referencia.
   - stt: Tipo de string.
   - long_str: Indica si el tamaño de la string es mayor de 255 caracteres
   - str: Valor de la string cuando es menor de 255 caracteres.
   - str_txt: Valor de la string cuando es mayor de 255 caracteres.
   - lat: Idioma.
   - lrl: Nivel de restricción que hay en la diversas fuente de UMLS (según la licencia).
4. **string_type (stt, descrp)**
   Recoge los diferentes tipos de string: preferida, en minúsculas, en singular, en plural, etc.
   - stt: Identifiador del tipo de string. Constituye la clave primaria.
   - descrp: Descripción del tipo de string.

En este caso simplemente se ha procedido a normalizar la tabla original dividiéndola en otras que se encuentran ya normalizadas. Solamente la información recogida en string_type no se encontraba de forma tabulada en el original, siendo necesario obtenerla de la documentación.
Una vez implementada la base de datos se procedió a su carga desde los ficheros originales. Para llevar a cabo este proceso se recurrió a la definición de un conjunto de pipelines con PowerBuilderâ, que conectándose mediante ODBC a los ficheros originales extraen los datos y los insertan en la base de datos destino. Ello implica en primer lugar identificar cuáles son las fuentes más fiables para cada elemento de información, dada la redundancia existente en el origen. Posteriormente hay que definir las consultas pertinentes para extraer los datos.

El proceso de carga no ha estado exento de dificultades técnicas. Se observó por ejemplo que había registros que, sin motivo aparente, no se transferían. Analizando de modo exhaustivo el contenido, se observó que en algunas entradas de los ficheros se encontraba el carácter comillas dobles (“). Este carácter no era correctamente interpretado por la herramienta, que desestimaba la entrada. En consecuencia hubo que proceder a sustituir todas las comillas dobles existentes en los ficheros originales por comillas simples (‘), tras lo cual se resolvió el problema.

## RESULTADOS

Aunque contamos con un diseño completo de una base de datos para almacenar UMLS, hasta el momento no hemos procedido todavía a la carga total de la misma, por lo que es previsible que sea preciso realizar todavía alguna modificación.

No obstante, lo que consideramos la porción principal ya está cargada, esto es, lo que hace referencia a los conceptos, términos, strings, palabras utilizadas en cada idioma, fuentes (clasificaciones originales de dónde se han tomado los conceptos) y tipo semántico. Quedan todavía por cargar los datos referentes a las distintas relaciones consideradas entre los conceptos (jerarquía, co-ocurrencia, contextos, etc.).

Como en la base de datos están declaradas las correspondientes restricciones de integridad, el intento de cargar un elemento duplicado o una clave foránea sin correspondencia genera el correspondiente error. Cabe destacar dos aspectos interesantes. De un lado, que la estructura original representa tener un alto índice de redundancia, como se muestra en la tabla 1 respecto al contenido de del fichero probablemente más importante, MRCON.

| Elemento | Entradas | Redundancias | Porcentaje |
|---|---|---|---|
| Concepto | 2.072.040 | 1.200.456 | 57,9 % |
| Término | 2.072.040 | 327.975 | 15,8 % |

**Tabla 1.- Redundancias**

Por otra parte, y aparte de la redundancia, se encuentra también un considerable número de duplicados, es decir, entradas realmente repetidas que no debían existir, como se muestra en la tabla 2.

| Elemento | Fichero origen | Entradas | Duplicados | Porcentaje |
|---|---|---|---|---|
| Concepto | MRSTY | 1.040.826 | 1.200.456 | 57,9 % |
| Término | MRCON | 2.072.040 | 17.110 | 0,8 % |
| String | MRCON | 2.072.040 | 8.451 | 0,4 % |
| Inglés | MRXW.ENG | 7.728.550 | 16.249 | 0,2 % |
| Francés | MRXW.FRE | 84.748 | 330 | 0,4 % |
| Alemán | MRXW.GER | 177.691 | 420 | 0,2 % |
| Español | MRXW.SPA | 180.485 | 3.947 | 2,2 % |
| Italiano | MRXW.ITA | 57.568 | 272 | 0,5 % |

**Tabla 2.- Duplicados**

Estos resultados implican que, de construir una base de datos fiel a la estructura de ficheros proporcionada por la NLM, la búsqueda de un concepto utilizando como clave una palabra que supuestamente esté contenida en alguna string relacionada, podría arrojar un enorme número de posibilidades, la mayoría de ellas repetidas, por el efecto multiplicador de la redundancia y los duplicados a cada nivel. Es decir, que sería poco eficiente y, lo que es peor, muy poco preciso.

Otro tipo de errores detectado hace referencia a conceptos vacíos de contenido. Se han encontrado 941 strings con contenido nulo, y 142 conceptos que están ligados únicamente a strings con contenido nulo.

Los primeros resultados obtenidos con la utilización de la base de datos son realmente alentadores. A falta de una cuantificación más precisa por el momento, sí podemos constatar su eficacia en la recuperación de conceptos relacionados con un conjunto de palabras en cualquiera de los idiomas soportados, con una respuesta casi inmediata.

## CONCLUSIONES

UMLS constituye seguramente la iniciativa más importante en el ámbito de la terminología médica. El enorme número de conceptos y expresiones de los mismos recogidos lo cualifican como marco de referencia terminológica en medicina.

No obstante, el sistema tal y como es entregado por la NLM adolece de una serie de inconvenientes que, unido al enorme volumen de datos a manejar, dificultan su utilización directa:

- Estructura escasamente conceptualizada, con información de diversa índole muy mezclada.
- Gran índice de redundancia
- Numerosos duplicados

El desarrollo de una base de datos correctamente diseñada y normalizada no sólo permite obviar todas estas dificultades, sino que hace entrever un gran número de utilidades futuras.

## REFERENCIAS


(1) SCOPE Web site. 2002. http://www.tecn.upf.es/scope/

(2) Lassila O, Swick RR. Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation. Lassila O, Swick RR, editors. 1999. http://www.w3.org/TR/REC-rdf-syntax/

(3) Brickley D, Guha RV. Resource Description Framework (RDF) Schema Specification 1.0. W3C Candidate Recommendation. Brickley D, Guha RV, editors. 27-3-2000. http://www.w3.org/TR/rdf-schema/

(4) Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Languaje System. Meth Inform Med 1993; 32:281-291.

(5) Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc 1994; 1(1):35-50.

(6) Cimino,J.J.; Hripcsak G; Johnson,S.B.; Friedman,C.; Fink,D.J.; Clayton,P.D.: UMLS as knowledge base - a rule-based expert system approach to controlled medical vocabulary management. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care. IEEE Computer Society Press, editor. 1990.

(7) Joubert M, Aymard S, Fieschi D, Volot F, Staccini P, Robert JJ et al. ARIANE: integration of information databases within a hospital intranet. Int J Med Inf 1998; 49(3):297-309.

(8) Leroy G, Chen H. Meeting medical terminology needs—the Ontology-Enhanced Medical Concept Mapper. IEEE Trans Inf Technol Biomed 2001; 5(4):261-270.

(9) Gu H, Perl Y, Geller J, Halper M, Liu LM, Cimino JJ. Representing the UMLS as an object-oriented database: modeling issues and advantages. J Am Med Inform Assoc 2000; 7(1):66-80.

(10) Pisanelli DM, Gangemi A, Steve G. An Ontological Analysis of the UMLS Metathesaurus. JAMIA 1998; 5(4):810-814.