ARTÍCULO ESPECIAL

Papeles Médicos 2001;10(3):136-143

Problemas terminológicos en medicina: ¿alguna novedad?*

J. Sempere
Unidad de Documentación Clínica y Admisión. Hospital Comarcal de Vinaròs

Correspondencia
Jordi Sempere Soler
Unidad de Documentación Clínica y Admisión
Hospital Comarcal de Vinaròs
Avgda. Gil d'Atrossillo, s/n
12500 Vinaròs
Tel.: 964 477 102
Fax: 964 400 736
E-mail: sempere_jor@gva.es

Introducción

El uso apropiado y preciso de un vocabulario específico sobre un área de conocimiento es crucial para la comunicación entre los especialistas en ese campo, y la medicina no es una excepción: bien se trate de observaciones recogidas sobre pacientes, de resultados de actuaciones preventivas o curativas sobre los mismos, o de conocimientos científicos que permitan analizar y explicar unas y otros, la información médica se transmite mayoritariamente a través de un lenguaje propio, basado en una terminología específica que designa los conceptos característicos de su área de interés. Los sistemas de recuperación de la información (SRI) médica dependen por tanto del uso y tratamiento de la terminología para cumplir adecuadamente sus funciones, y en este sentido los problemas terminológicos ocupan un lugar preferente dentro del campo de la Documentación médica, y han justificado en gran medida su aparición como ciencia aplicada.

Como documentalista, cuando uno revisa la bibliografía aparecida sobre el tema en los últimos años, lo hace a la vez con asombro y con desazón. La causa del asombro proviene del inusitado aumento del número de artículos que durante la última década, han aparecido en la literatura médica consagrados al tema del control de vocabulario, la indización y codificación de información médica y otros aspectos relacionados con problemas terminológicos.

En un análisis a simple vista, la mayor parte de artículos aparecen publicados en dos revistas que constituirían un improvisado núcleo de Bradford sobre el tema. Se trata de una revista americana, *Journal of the American Medical Informatics Association*, y de otra europea, alemana concretamente, *Methods of Information in Medicine*. Ambas revistas se dedican a eso que, fuera de nuestras fronteras, se ha dado en llamar Informática médica. Un reconocido experto en la materia, Enrico Coiera[1], señala irónicamente que "la Informática médica va tanto de ordenadores como la cardiología de estetoscopios". Nosotros no tenemos nada que objetar al respecto, únicamente añadiríamos que la Informática médica trata, en gran medida, de Documentación médica. Creemos, por lo demás, estar del mismo lado que Coiera cuando a continuación añade que "cualquier intento de implantar la informática fracasará estrepitosamente si la motivación es la aplicación de la informática en sí misma más que la solución de los problemas clínicos". Tampoco creemos casual que la revista *Methods...*, aparecida originalmente en 1957 bajo el título *Medizinische Dokumentation*, título que en el año 1961 cambió por el actual, sea la primera revista internacional sobre nuestra especialidad[2].

En cualquier caso, el tema que ahora nos preocupa, no es el de las denominaciones de nuestra área de conocimiento, por lo demás suficientemente aclarado en otros trabajos[2,3]. Lo que nos causa la desazón a la que aludía en un principio, son ciertas actitudes subyacentes en estos artículos que en general podríamos resumir en dos aspectos: en primer lugar, una ignorancia prácticamente absoluta de los antecedentes históricos de los problemas que se analizan, y en segundo lugar, y quizá en relación con lo anterior, un uso negligente e inapropiado de conceptos, que procedentes del dominio de la Documentación, tienen significados precisos y perfectamente acotados.

Efectivamente, incluso en los artículos de revisión o de síntesis[4-8] quizá más acentuadamente en unos autores que en otros, hay una exposición de los problemas terminológicos de tendencia marcadamente empirista, orientada a la necesidad práctica de establecer una terminología médica de uso internacional, dentro del contexto de una automatización del soporte de los documentos médicos en general, y de la historia clínica en particular. Si en algún caso se hace un análisis histórico, suele ser superficial, sin profundizar en las causas que, a lo largo del tiempo y aún en la actualidad, han impedido e impiden la consecución de una clasificación coherente y estable de las enfermedades, y en relación con ello, de una nomenclatura nosológica universalmente aceptada.

Por otra parte se habla indiscriminadamente de sistemas de codificación, de terminologías médicas o de vocabularios controlados, y se incluyen bajo estas denominaciones lo mismo la *Clasificación Internacional de Enfermedades* (CIE), que la *Standard Nomenclature of Medicine* (SNOMED) o un tesauro como los *Medical Subject Headings* (MeSH). Si nosotros tuviéramos que agrupar bajo una misma categoría elementos tan dispares hablaríamos en todo caso de lenguajes documentales o lenguajes de indización. También se manejan con mayor o

*Artículo basado en la ponencia titulada "Problemas y soluciones al tratamiento de la terminología" presentada en el VII Congreso Nacional de Documentación Médica. Valencia, junio de 2001

menor propiedad términos documentales como precoordinación, control de vocabulario, etc. sin situarlos en su contexto, o peor, se acuñan innecesariamente términos nuevos (por ejemplo, se utiliza el término "*granularity*" para referirse al grado de detalle o especificidad de un determinado vocabulario; como siempre en estos casos también hay epígonos españoles que comienzan a traducir sin ningún escrúpulo "*granularidad*").

Desde esta perspectiva, los problemas terminológicos se abordan como relativamente nuevos, del mismo modo que las soluciones que se proponen, que ajenas al sentido de aportaciones anteriores, se revisten de un carácter innovador. ¿Cómo si no entender que se despache una institución taxonómica como la CIE, simplemente porque no tiene el nivel de especificidad adecuado para servir de lenguaje de expresión en las historias clínicas? ¿Cómo es posible que se pretenda evaluar una clasificación como si fuera una terminología? Y más cuando en el prólogo de la propia CIE ya se delimitan claramente las diferencias entre ambas[9].

Trataré a continuación de analizar, en qué medida son nuevos los problemas terminológicos que se plantean; a continuación expondré las soluciones que se proponen o investigan actualmente, y finalmente intentaré extraer alguna conclusión del análisis efectuado.

¿Los problemas son nuevos?

Un aumento del número de trabajos que abordan los problemas terminológicos médicos lleva a pensar que como mínimo, hay elementos de juicio nuevos que influyen en una concepción diferente de los problemas.

Después de una lectura detenida de la bibliografía, cualquiera puede comprobar que los problemas son los mismos de siempre, y que la única circunstancia que quizá ha cambiado, o está en vías de hacerlo, es el soporte de la información. Es decir, se echa en falta y se plantea la necesidad de una terminología normalizada y universalmente aceptada (es decir, lo que denominaríamos una nomenclatura), en relación con una informatización del soporte, no ya sólo de las publicaciones medicocientíficas, sino también y muy particularmente, de los documentos primarios medicoasistenciales, es decir, de la historia clínica.

Ciertamente, de una informatización adecuada de la historia clínica cabe esperar no sólo un registro más o menos ordenado de datos, sino también y de manera especial, una recuperación adecuada de la información clínica que permita atender los diferentes usos previsibles de la misma (primariamente asistenciales, pero también con fines de investigación, docencia, planificación y gestión de recursos, evaluación de la calidad asistencial, etc.).

Y en ello influiría decisivamente disponer de una nomenclatura que pudiera ser utilizada no sólo dentro de una misma institución, sino también a escala suprainstitucional, y que además fuera compartida por todas las fuentes de información médicas (fueran éstas científicas, asistenciales o sanitarias)[10-12].

No obstante, una cosa es el deseo y otra la realidad. Un análisis histórico de los hechos revela que la aspiración no es ni mucho menos nueva, pero que tampoco se dan las circunstancias que indiquen que los problemas de fondo que obstaculizan su consecución hayan variado[13,14]. En efecto, es muy difícil conseguir una nomenclatura si no está respaldada por una clasificación consistente de los saberes nosológicos, y ésta, que sepamos, todavía no existe.

Desde que a finales del siglo XVII, la patología abandona los postulados esencialistas de la taxonomía galénica y pasan a adoptarse los criterios metodológicos de la ciencia moderna (proceso que suele personalizarse en la figura de Thomas Sydenham), las enfermedades se convierten en abstracciones o categorías conceptuales obtenidas por razonamiento inductivo a partir de la observación directa de las formas de enfermar. La renovación de la patología propiciada por el programa de Sydenham pronto dará lugar a una proliferación de intentos sistematizadores del saber médico, y de manera consecuente, a la necesidad de una terminología médica de uso universal. A lo largo de un complejo proceso histórico que arranca en la segunda mitad del siglo XVIII, la nosología moderna ha ido construyendo un modelo nosotáxico basado en los conocimientos aportados por la etiología, la anatomía patológica, la fisiopatología, la patocronia y la semiología. Pero ninguna de estas disciplinas por separado ha dado una explicación integral del proceso de enfermar.

Por las propias características del método científico, los mismos conceptos de enfermedad están sujetos a revisión continua, no ya sólo por la aparición de nuevos medios técnicos de investigación, sino también por factores socioculturales que afectan tanto a la comunidad científica que las define, como a la propia sociedad en la que aquélla se inscribe. En este sentido, como señala Barona[15], la sistematización de las enfermedades mentales "es un caso paradigmático de las dificultades que ha de afrontar la ciencia médica moderna para establecer categorías diagnósticas científicamente consistentes".

Por otra parte, en la práctica clínica es frecuente el uso de clasificaciones que no siempre siguen los criterios de ordenación teóricos propios de una clasificación de intención global como la CIE; la utilidad de una clasificación clínica viene dada por su capacidad de agrupar consistentemente y sin sesgos, casos individuales según criterios terapéuticos y pronósticos[14]. Es frecuente, además, encontrar varias clasificaciones que abordan simultáneamente un mismo problema clínico, y la elección final de una o de otra dependerá de la validez demostrada por cada una en su uso por diferentes grupos de médicos prácticos. Esto determina que, incluso en cuestiones aparentemente resueltas desde un punto de vista nosológico, como puede ser el caso de las fracturas, todavía haya disensiones respecto a la manera de clasificarlas y denominarlas[16].

Junto a estos problemas de orden conceptual, que resultan en una provisionalidad permanente no sólo de los conceptos médicos, sino también de los criterios que permiten su ordenación, y que por tanto, obligan a una necesaria clasificación multiaxial de los mismos, existen otros problemas relacionados con la forma en que se expresan dichos conceptos, es decir, con el lenguaje.

El lenguaje médico, al igual que todo lenguaje científico, tiene como objetivo referirse con precisión a los conceptos propios de su área de conocimiento, al tiempo que pretende servir de vehículo de comunicación entre los miembros de la comunidad médica internacional. Para ello se sirve de una terminolo-

gía específica que en general utiliza vocablos procedentes de las lenguas clásicas, o crea neologismos basados en raíces, prefijos o sufijos de origen grecolatino, de modo que los términos tienen una base de representación similar en las lenguas habladas por los médicos de diferentes países.

A pesar de ello, el lenguaje médico no siempre se adapta a un criterio lógico uniforme, y además participa de las imprecisiones y ambigüedades propias del lenguaje natural[17]. La ambigüedad deriva de circunstancias tales como la variación diacrónica de los significados de los términos, de la utilización de epónimos y acrónimos que tienen muchas veces un ámbito de uso limitado e incluso puramente coyuntural, o de la presencia de fenómenos semánticos como la homonimia, la polisemia y la sinonimia (o más genéricamente la paráfrasis).

A estos problemas cabría añadir el de los barbarismos o faltas de forma y empleo correcto de las palabras, entre los que destaca el uso innecesario de términos procedentes de otros idiomas por traducción incorrecta (habitualmente anglicismos) o por asimilación inadecuada a través de otras lenguas (habitualmente el inglés) de neologismos de base grecolatina[18].

¿Son nuevas las soluciones?

Del mismo modo que los problemas no eran del todo nuevos, tampoco lo son las soluciones, o al menos, no de una manera radical.

Los sistemas de recuperación de la información han sido la respuesta metodológica y técnica de la Documentación a los problemas de transmisión de la información científica[2]. Un elemento crucial en todo SRI es el lenguaje documental utilizado, en la medida en que controla la variabilidad del lenguaje natural existente en los documentos primarios, y posibilita un almacenamiento y una recuperación de la información acorde con los objetivos del sistema[19,20].

En la síntesis que Lancaster[19] publicó a mediados de los ochenta sobre el control de vocabulario en los sistemas de información, señalaba dos grandes líneas de desarrollo futuro de los mismos: una primera, relacionada con la recuperación automática de información directamente a partir del lenguaje natural; la segunda apuntaba hacia la posibilidad de que los lenguajes documentales de diversos SRI temáticamente similares, fueran compatibles entre sí e interconvertibles; un desarrollo subsidiario de esta segunda tendencia perseguiría la compatibilidad no ya de los lenguajes documentales, sino de la terminología en general, mediante la elaboración de grandes bancos de datos terminológicos. Las soluciones al tratamiento de la terminología en la bibliografía comentada encajan perfectamente en las previsiones de Lancaster.

En realidad estas tendencias de desarrollo vienen a evidenciar la tensión existente entre dos fuerzas contrapuestas en todo sistema de información[7,8,21]: de un lado, la necesidad de expresividad representada por el lenguaje natural, que permite un desarrollo y crecimiento libre de los términos y una máxima especificidad en la recuperación, deslucida de alguna manera por las ambigüedades, incorrecciones e imprecisiones que le son propias y que actúan dificultando una recuperación consistente de la información; de otro lado, la necesidad de normalización que impone el lenguaje documental, que libera de ambigüedades e incorrecciones el lenguaje al tiempo que lo encorseta y empobrece, y que permite una máxima consistencia y exhaustividad en la recuperación de información aun a costa de cierta pérdida de especificidad.

En la historia clínica informatizada posiblemente este conflicto adquiere una mayor intensidad, dado que, por un lado, la informatización suele requerir en sí misma una mayor estructuración de la entrada de información que la que exige una historia clínica manual; y por otro lado, la historia clínica es quizá, de todos los documentos primarios médicos, el que un mayor grado de expresividad requiere, en consonancia con el carácter personal e individual de su contenido, con la contingencia de los fenómenos que se describen y con la provisionalidad de las actuaciones que se practican, dependientes por necesidad de resultados igualmente contingentes.

Esquemáticamente, la solución buscada es la de un SRI (particularmente la historia clínica) en el cual la entrada de información se produce en lenguaje natural y con el grado de especificidad deseado, y tanto el procesamiento como la recuperación de información son procesos automatizados. La recuperación además debe adaptarse a los diferentes usos de que es objeto la historia clínica. Para ello, es preciso que el lenguaje documental del sistema permita tanta especificidad como la que pueda expresarse a la entrada, al tiempo que garantice una ordenación multiaxial de los conceptos, de manera que sea posible su recuperación desde diversos puntos de vista. No es sorprendente, a la vista de lo expuesto, que algunos autores expresen los requerimientos de un tal lenguaje en forma de desiderátum[5].

Las posibilidades de introducción de datos en sistemas informatizados son en general tres, cada una con peculiaridades diferentes de cara a la recuperación de información:

1. Texto libre
2. Listas predeterminadas, que a su vez, pueden ser:
   2.1. listas precoordinadas o enumerativas, o
   2.2. listas postcoordinadas o compositivas, y
3. listas mixtas, que permiten la introducción libre de términos al tiempo que la selección de expresiones de una lista.

El texto libre permite una máxima expresividad pero unas posibilidades de recuperación muy limitadas, aunque útiles en la recuperación de conceptos muy específicos. En general, la indización mediante el tratamiento automatizado del lenguaje natural (*natural language processing*) sólo ha mostrado alguna efectividad en dominios concretos y restringidos (véanse por ejemplo, los trabajos de Friedman *et al.*[22] sobre el análisis de informes radiológicos o el proyecto MENELAS sobre informes de alta de pacientes con enfermedades coronarias[23]). Las dificultades de los ordenadores para comprender la información contextual o implícita, y para depurar adecuadamente las ambigüedades semánticas y/o sintácticas de un texto, son los principales inconvenientes con los que tropieza.

Las listas predeterminadas, y especialmente las precoordinadas, tienen el inconveniente de que limitan enormemente la expresividad, y pueden empobrecer el lenguaje y la información que se introduce: el usuario debe ceñirse necesariamente a un repertorio finito de expresiones que no siempre estará de acuer-

do con sus necesidades; las listas postcoordinadas, al permitir la composición de conceptos complejos a partir de otros más simples aumentan las posibilidades de expresión al tiempo que permiten una mayor economía de mantenimiento: considérense por ejemplo[7] los términos que sería necesario añadir para describir los diferentes tipos de quemaduras en una lista precoordinada: considerando que puede haber dos tipos básicos, térmicas y químicas, 200 localizaciones anatómicas posibles, tres grados de profundidad diferentes más el no especificado, y tres grados diferentes de extensión más el no especificado, tendríamos $2 \times 200 \times 4 \times 4 = 6.400$ expresiones diferentes sólo para describir las quemaduras, y ello sin tener en cuenta otras características de las mismas, como la presencia o no de infección, etc. Las listas precoordinadas si quieren garantizar una exhaustividad en la cobertura crecen en progresión geométrica y presentan un rendimiento bajo: en el ejemplo de las quemaduras, aun suponiendo que estuviéramos en un centro de quemados, habría combinaciones que probablemente no se emplearían nunca. En general, si se opta por una lista predeterminada, y especialmente si se elige un enfoque precoordinado lo habitual es que se trate de un sistema mixto, que permita la introducción de expresiones por parte del usuario, y que irá enriqueciéndose y ampliando su cobertura sólo en la medida en que el propio usuario lo necesite: aun así siempre existirá la tendencia a que haya un pequeño número de expresiones de frecuencia de utilización media o alta, y un gran número de expresiones de uso esporádico[24,25], lo cual obligará inevitablemente a llevar un mantenimiento de la misma que impida su crecimiento fuera de márgenes razonables por simple adición de expresiones de poco uso.

Las listas predeterminadas, sean del tipo que sean, pueden favorecer la uniformización de las expresiones introducidas por los usuarios, por la tendencia de éstos a escoger términos de una lista preestablecida, aunque no se adapten completamente a la idea que querían expresar, a cambio del esfuerzo relativo que supone introducir una expresión nueva.

Como ventajas las listas predeterminadas presentan la posibilidad de controlar la variabilidad innecesaria y las incorrecciones, siempre que exista una validación de las expresiones introducidas; también pueden permitir una recogida exhaustiva de información necesaria para clasificar/codificar posteriormente con el máximo grado de detalle, especialmente si la entrada de datos está muy estructurada, aunque esto puede tener un efecto disuasorio sobre el usuario que puede dejar de introducir la información por el esfuerzo que conlleva; facilitan en principio la recuperación, aunque la versatilidad de la misma dependerá de la riqueza de los lenguajes documentales utilizados y de las relaciones semánticas que se establezcan entre los términos o expresiones de la lista, es decir, del control de vocabulario existente.

En nuestro medio son habituales las listas mixtas precoordinadas de diagnósticos, procedimientos o problemas, asociadas o no a códigos de la CIE u otros: muchos sistemas presentan además la posibilidad de codificación automática por reconocimiento de expresiones idénticas codificadas previamente[25-27]. Estos sistemas facilitan en gran medida la engorrosa tarea de codificación de expresiones que tienen frecuencias medias y altas de uso, pero no suelen alcanzar una cobertura completa dado que, como se ha indicado antes, siempre hay un gran número de expresiones de frecuencia de aparición muy baja que no suelen codificarse automáticamente. En general favorecen la fiabilidad de la codificación al automatizar el proceso, pero no necesariamente aumentan la exactitud de la misma, o incluso la pueden perjudicar (así, por ejemplo, si se asigna el código de la CIE-9-MC 774.6 de la categoría "Ictericia neonatal", a una expresión que sólo indique "Ictericia" aunque implícitamente se sepa que corresponde a un recién nacido, se estará favoreciendo un sesgo de clasificación, dado que el sistema tenderá a codificar todas las ictericias no especificadas como neonatales). La exactitud de la codificación en estos sistemas pasa por asegurar que la expresión en lenguaje natural contiene una perfecta representación de las características de la categoría representada por el código, y esto inevitablemente, conduce a expresiones más complejas y variables que son más difíciles de codificar automáticamente. Por otra parte, no siempre el grado de coordinación existente en la expresión permite su correlación unívoca con un código, por lo que será necesario descoordinar o descomponer expresiones; y a la inversa, a veces será necesario juntar dos expresiones diferentes, que una vez coordinadas corresponden a una clase o código más adecuado. Para evitar sesgos, conviene igualmente que la persona que indiza o introduce expresiones no esté implicada en el proceso de codificación posterior (aunque sí deba necesariamente conocer su mecánica y los criterios que se siguen): podría ocurrir que empobreciera más o menos conscientemente determinadas expresiones que supiera ya codificadas para asegurarse de que fueran codificadas automáticamente. Todas estas cuestiones limitan el alcance de los sistemas de codificación automática basados en listas de expresiones previamente codificadas, y constituyen aspectos a investigar si se plantea una mejora de los mismos, pero a pesar de todo pueden ser herramientas útiles si se manejan adecuadamente.

Por lo dicho anteriormente es razonable pensar que una entrada compositiva o postcoordinada es la que mejor se adapta a las exigencias de una historia clínica informatizada; al fin y al cabo la composición es un mecanismo natural del lenguaje para crear términos más complejos: la posibilidad de crear diagnósticos complejos a demanda, por combinación a partir de un repertorio de términos simples es mucho más funcional que un enfoque precoordinado, en el cual sería preciso enumerar todas las combinaciones posibles de antemano. Tal como observaba Grémy[28] a finales de los 80 y se ha confirmado posteriormente con creces, no es casual que una clasificación facetada como la SNOMED, de difícil uso en SRI manuales, haya conocido un desarrollo espectacular paralelo a la mejora de las técnicas informáticas.

No obstante, el enfoque postcoordinado o compositivo también plantea dificultades, no sólo como sistema de introducción de datos sino también con vistas al almacenamiento y la recuperación de información.

Por lo que respecta a la introducción de datos hay que tener en cuenta que no todas las combinaciones posibles son válidas ni tienen sentido, y es necesario establecer ciertas restricciones. Por otra parte, las posibilidades de combinación pueden permitir en algún caso crear diversas combinaciones para definir un mismo concepto, y es indispensable que todas ellas

sean reconocidas por el sistema como equivalentes. Una mera yuxtaposición de términos, como la que se daba en antiguas versiones de la SNOMED no siempre tenía un sentido preciso, y además era necesario en ocasiones establecer relaciones entre elementos de una misma faceta, que en principio no eran coordinables entre sí.

Por lo que respecta al almacenamiento y la recuperación, no siempre es fácil para un ordenador interpretar una expresión compuesta del lenguaje natural, incluso aunque disponga de los términos simples que permiten componerla[7,8]. Por ejemplo para una computadora, puede ser difícil entender que "nefrectomía izquierda" se refiere a "exéresis del riñón izquierdo" y no a "exéresis izquierda del riñón". O que "hemorragia digestiva" equivaldría a "hemorragia" "del" "tubo digestivo", y que el nexo "del" debería interpretarse como "hemorragia" (localizada en) "tubo digestivo". Este tipo de deducciones pueden ser difíciles incluso para una persona que no tenga un conocimiento previo del campo conceptual en cuestión. Por otro lado, aunque teóricamente el tubo digestivo está comprendido entre la boca y el ano, ningún clínico incluiría una hemorragia gingival o faríngea dentro del concepto de hemorragia digestiva. Resulta por tanto difícil expresar formalmente la información implícita que a veces existe en términos complejos.

Todas estas reflexiones conducen a pensar que en un sistema postcoordinado deben existir términos con cierto grado de precoordinación, pero también a que es necesario definir relaciones precisas entre términos y jerarquizar términos y relaciones, y crear reglas específicas de composición de términos, en fin, llegar a cierto grado de formalización que permita un procesamiento y una clasificación adecuada de los diferentes conceptos, si se piensa en su recuperación automática[8,11].

Las relaciones entre conceptos deben quedar explicitadas; así, por ejemplo, los conceptos que subyacen en el término apendicitis aguda podrían expresarse formalmente como aparecen en la Figura 1 (representamos arbitrariamente los conceptos entre corchetes y las relaciones entre paréntesis; obsérvese que las relaciones son direccionales y que podrían serlo en un sentido o en ambos).

Las relaciones deben ser restrictivas en el sentido de que sólo pueden ser válidas con cierto tipo de conceptos: por ejemplo, puede señalarse que la relación (localizada en) sólo es válida para vincular conceptos morfológicos o funcionales con localizaciones anatómicas, pero para vincular conceptos de la categoría procedimientos con la categoría localización anatómica debe usarse la relación (actúa sobre).

También deben existir reglas de composición de conceptos, de modo que los conceptos ligados por la relación "localizada en" se ajusten a lo expresado en el diagrama de la Figura 2.

A partir de aquí, si hay una jerarquización adecuada de los conceptos de modo que inflamación es un concepto específico de lesión, y apéndice ileocecal es más específico que localización anatómica, la relación (localizada en) para unir los dos conceptos de apendicitis, inflamación y apéndice ileocecal, es correcta. La misma jerarquía, una vez establecidas las relaciones entre los conceptos más genéricos, permitiría recuperar apendicitis tanto en una búsqueda de lesiones del apéndice como en otra sobre procesos inflamatorios del aparato digestivo.

Figura 1. **Relaciones entre conceptos**

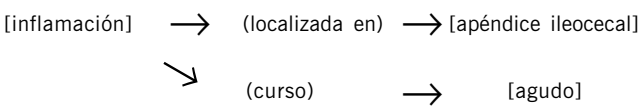
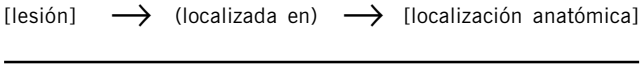


Figura 2. **Regla de composición de conceptos**



En definitiva, lo que se pretende es un modelo o representación conceptual (*knowledge representation*) que sirva de base a un programa informático. Tal representación toma como vehículo de expresión determinados lenguajes formales desarrollados a partir de los años 70 por investigadores procedentes del campo de la Inteligencia artificial: de estos lenguajes los más utilizados han sido las lógicas descriptivas (*description logics*) y los diagramas conceptuales (*conceptual graphs*). La base de una representación formal de este tipo radica sobre tres factores:

1. Una tipificación y jerarquización de los conceptos.
2. Una tipificación de las relaciones entre los conceptos.
3. Unas reglas que permitan combinar conceptos y relaciones para componer conceptos complejos.

Al conjunto de estos tres elementos es a lo que se suele llamar estructura, y a esta misma estructura, dotada de un contenido concreto es a lo que en la bibliografía se suele denominar con el pretencioso término de *ontología* (*ontology*). La tal *ontología* sería, por así decir, el modelo o gramática formal de una terminología[8,11].

En la actualidad hay en marcha tres proyectos de investigación, que cuentan con respaldo económico millonario, para llevar adelante lo que habrían de ser los lenguajes documentales médicos de referencia para uso en sistemas informáticos:

1. El UMLS o *Unified Medical Language System*, de la *National Library of Medicine*.
2. La SNOMED-CT (SNOMED-*Clinical Terms*), que cuenta con el patrocinio fundamental del *College of American Pathologists* (CAP) y el *National Health Service* (NHS) británico.
3. El proyecto GALEN y su continuación GALEN-IN-USE, que cuenta con el apoyo de la Unión Europea.

Los dos primeros tienen en común que toman como base lenguajes documentales existentes previamente (MeSH y SNOMED respectivamente), en tanto que GALEN, más modesto económicamente, nace como necesidad de desarrollo de un sistema de historia clínica informatizada, el PEN&PAD, desa-

rrollado en la Universidad de Manchester. Un análisis pormenorizado de cada uno sería objeto de un artículo independiente, por lo que nos limitaremos a señalar sus características fundamentales.

Unified Medical Language System (UMLS)

Se trata de un proyecto auspiciado por la *National Library of Medicine* (NLM) en colaboración con grupos de investigadores de diversas universidades norteamericanas[29,30]. Comienza en 1986 y en la actualidad todavía está en fase de desarrollo.

Una particularidad de UMLS respecto a otros proyectos es que no consideró inicialmente, debido a su gran envergadura y complejidad, elaborar una clasificación de conceptos médicos (en la línea de las *ontologías*) y optó por establecer correlaciones entre términos de diferentes vocabularios representando conceptos similares. Esta idea fue objeto de gran controversia al inicio del proyecto, prevaleciendo los criterios de la NLM frente a la propuesta de muchos colaboradores, que consideraban esencial que el UMLS tuviera una adecuada cobertura del vocabulario específico utilizado en las historias clínicas. En general todo el proyecto ha sido el fruto de esta dialéctica continua[31]. Con el tiempo, no obstante, el UMLS ha ido incorporando fuentes de terminología más propiamente clínica.

El UMLS se estructura en torno a tres componentes, denominados fuentes conceptuales (*knowledge sources*), las cuales se combinan con un conjunto de programas de acceso a las mismas. Las fuentes son las siguientes:

1. *Metathesaurus*
2. *Semantic Network*
3. *Specialist Lexicon*

El *Metathesaurus* es una unión controlada de diversas fuentes terminológicas, entre las que se encuentran el propio tesauro MeSH (traducido a diversas lenguas), diversas clasificaciones como la CIE o la Clasificación Internacional de Atención Primaria de la WONCA, diferentes ediciones de la SNOMED, vocabularios procedentes de sistemas de historias clínicas informatizadas, incluso sistemas de ayuda al diagnóstico como AI/RHEUM, y bases de datos factuales sobre fármacos, tóxicos, proteínas, genes, etc. En la versión de 2001, el UMLS agrupa 60 fuentes terminológicas diferentes, cuenta con unos 800.000 conceptos y 1,9 millones de términos en diferentes idiomas.

El UMLS no necesariamente vacía todos los términos de las fuentes terminológicas de las que se alimenta. Se organiza de acuerdo con el concepto y mantiene la vinculación del mismo con los diferentes términos que lo denominan y sus variantes léxicas. El origen de cada término está rigurosamente preservado, e incluso se conservan las definiciones y las relaciones semánticas del término en su fuente original. No obstante, algunas de estas relaciones, pasan a formar parte de la red semántica del propio UMLS.

La red semántica (*Semantic network*) es la que confiere consistencia a una diversidad de conceptos de diferentes fuentes con sus propias ordenaciones y estructuras. Se articula sobre la definición de unas categorías semánticas elementales (*semantic types*) de las cuales hay 134 en la última versión, y que actúan como los nudos de la red, y de una serie de relaciones entre dichas categorías (los hilos de la red), de las que hay definidas 54 tipos. Ejemplos de categorías semánticas son: organismos, estructuras anatómicas, funciones biológicas, productos químicos, etc.). Cada concepto del *Metathesaurus* puede ser asignado a varias categorías diferentes, siguiendo métodos algorítmicos y manuales escrupulosamente validados. La relación primaria de la red semántica es la paradigmática pura, esto es, "tipo de" o "clase de" (en inglés "*is a"*). Otro tipo de relaciones esenciales, pero no jerárquicas, son: la espacial, la funcional, la temporal, la conceptual y la física. Las relaciones, en lo posible, se refieren a categorías conceptuales genéricas: de esta manera se favorece un almacenamiento eficiente de la información, dado que a través de la jerarquía, el ordenador puede deducir propiedades para términos específicos a partir de las que poseen los genéricos de los que nacen. Por supuesto, con las debidas restricciones a que pueda haber lugar.

El *Specialist Lexicon*[32], se incorporó a partir del año 94 a las fuentes conceptuales del UMLS. Recoge un conjunto de lexemas (que pueden ser palabras simples, compuestas o partículas, según el caso), referidos al campo biomédico y del idioma inglés en general, junto con información morfológica, sintáctica y ortográfica asociada a los mismos. La información semántica sobre dichos lexemas se puede encontrar en la mayoría de los casos en el *Metathesaurus*. Actualmente el *Specialist* contiene 140.000 entradas. Se asocia a programas que permiten identificar o generar variantes léxicas de un mismo término y que resultan muy útiles para el tratamiento automático del lenguaje natural. También permiten la construcción de índices de acceso al contenido del *Metathesaurus*, de manera que es posible introducir por ejemplo la palabra "*heart*" y recuperar consistentemente términos formalmente tan diferentes pero conceptualmente relacionados como "*American Heart Association*", "*Cardiac volume*" o "*Coronary artery by-pass*".

SNOMED-CT (SNOMED Clinical Terms)

Desde la aparición de la *Standard Nomenclature of Pathology* (SNOP) en 1965, con cuatro ejes de indización, ésta ha ido experimentando sucesivos desarrollos[33-35]: en 1977 aparece como SNOMED incorporando dos nuevos ejes de enfermedades y procedimientos; en 1979 incorpora un séptimo eje de ocupación y pasa a denominarse SNOMED II. En 1993 aparece una nueva edición, bajo el nombre de SNOMED *International*, en la que participa también la *American Veterinary Medical Association*, y que pasa a tener 11 ejes o facetas, por división del eje etiológico en cuatro diferentes relativos a organismos vivos, compuestos químicos y biológicos, agentes físicos y dispositivos médicos, y contexto social, más un último eje de modificadores genéricos destinados a vincular de manera específica unos conceptos con otros. La SNOMED *International* tiene revisiones anuales desde el 94 hasta el 97. A partir del 96, comienza a desarrollarse un nuevo proyecto, SNOMED-RT (SNOMED *Reference Terminology*) auspiciado por el CAP, una compañía de seguro privado americana, Kaiser Permanente, y la Clínica Mayo, que incorpora el uso de una lógica de descripción (K-Rep) para definir los conceptos y las relaciones entre ellos, de modo que de manera automática puedan reconocerse diferentes expresiones de significado equivalente. El código deja de tener un sentido jerárquico y un mismo concepto puede formar parte de más de una jerarquía.

El proyecto actual SNOMED-CT, cuya primera versión tiene prevista su aparición a finales de 2001, persigue la convergencia de la SNOMED-RT y la versión más desarrollada de los códigos Read, los *Clinical Terms* del NHS[36].

Los códigos Read (*Read Codes*)[37,38] nacen a comienzos de los años 80 por iniciativa de un médico generalista británico, James Read, con la finalidad práctica de llevar un seguimiento sistemático de grupos de pacientes con enfermedades crónicas o problemas de salud específicos. El sistema de codificación va adquiriendo desarrollo, y en el año 90 es comprado por el NHS e impulsado como una clasificación para uso en todos los niveles asistenciales, evolucionando desde una estructura jerárquica clásica a un planteamiento multiaxial y "tesaurizado".

Proyecto GALEN (Generalised Architecture for Languages Encyclopaedias and Nomenclatures in medicine)

Se trata de un proyecto patrocinado por diferentes entidades y organismos europeos[39].

GALEN aspira a crear un modelo conceptual formalizado del dominio médico en el sentido de las *ontologías* mencionadas anteriormente. A fin de poder establecer un vínculo entre la entrada de datos por parte del usuario y la traducción al lenguaje formal, se propone la creación de un distribuidor o servidor terminológico, llamado TeS (*Terminology Server*), que pueda ser utilizado por diversas aplicaciones.

Internamente, el servidor se compone de tres módulos: 1. un módulo conceptual (*Concept Module*, CM) que contiene el modelo conceptual de referencia (*Concept Reference* CORE *model*) , expresado en un lenguaje formal (GRAIL, una variedad de lógica descriptiva); 2. Un módulo multilingüe (*Multi-lingual Module*, MM) para representar los conceptos del CORE mediante términos de diferentes lenguas europeas (entre las que no figura el castellano) y 3. Un módulo de conversión de los conceptos del CORE a los sistemas de clasificación habituales (*Code Conversion Module*, CCM)[40].

Conclusiones

Las aportaciones científicas de la Documentación están enmascaradas por el efecto apabullante del desarrollo de las técnicas informáticas.

En los últimos años estamos asistiendo a un proceso inexorable de informatización del soporte de la información médica. Este hecho ha propiciado una preocupación renovada por la creación de herramientas terminológicas adaptadas a las necesidades del procesamiento y recuperación automáticos de la información médica.

No hay indicios que permitan predecir la consecución de una nomenclatura médica a medio y largo plazo. Los proyectos de desarrollo de una terminología médica de referencia parecen estar llegando a un punto de saturación, dado que han adquirido un tamaño y complejidad crecientes que complicarán su mantenimiento y actualización futuros, sin que hasta la fecha se haya podido demostrar que puedan alcanzar una cobertura terminológica completa de las fuentes de información biomédicas[1].

A pesar de todo es muy posible que a corto plazo tengamos que enfrentarnos a utilizar aplicaciones informáticas que se sirvan de este tipo de productos terminológicos. De hecho, diversos productos de la NLM, entre ellos uno tan conocido como PubMed, se basan en fuentes proporcionadas por el ULMS.

Como documentalistas estamos obligados a conocer de primera mano los nuevos productos, y examinar sus posibilidades de adaptación a nuestras necesidades particulares, entre las que figura de manera destacada la incorporación a dichas herramientas de las lenguas habladas en nuestro medio.

Bibliografía

1. Coiera E. Recent advances: Medical informatics. *Br Med J* [edición electrónica] 1995;310(6991):1381-7 [citado 14 Sep 2001]. Disponible en URL: http//www.bmj.com/cgi/content/full/310/6991/1381.
2. Terrada ML. *La Documentación médica como disciplina*. Valencia: Centro de Documentación e Informática Biomédica, 1983.
3. Peris Bonet R. Documentación médica hospitalaria en España. Algunas reflexiones desde Valencia. *Papeles Méd* 1998;7(1):18-24.
4. Cimino JJ. Review paper: coding systems in health care. *Meth Inf Med* 1996;35(4-5):273-84.
5. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Meth Inf Med* 1998; 37(4-5):394-403.
6. Chute CG. Clinical classification and terminology: some history and current observations. *J Am Med Inform Assoc* 2000;7(3): 298-303.
7. Rector AL. Clinical terminology: why is it so hard? *Meth Inf Med* 1999;38:239-52.
8. Zweigenbaum P. Encoder l'information médicale: des terminologies aux systèmes de représentation des connaissances. *Innov Stratég Inf Santé* 1999;(2-3):27-47. Disponible en URL: http//www.biomath.jussieu.fr/~pz/Publications/biblio-pierre-pardate. [Citado 14 Sep 2001].
9. *Manual de la Clasificación Estadística Internacional de Enfermedades, Traumatismos y Causas de Defunción*. 9ª Revisión. Washington: Organización Panamericana de la Salud, Organización Mundial de la Salud, 1978.
10. Barnett GO, Jenders RA, Chueh HC. The computer-based clinical record. Where do we stand? *Ann Intern Med* 1993;119(10): 1046-8.
11. Evans DA, Cimino JJ, Hersh WR, Huff SM and the Canon Group. Toward a medical-concept representation language. *J Am Med Inform Assoc* 1994;1(3):207-17.
12. Board of Directors of the American Medical Informatics Association. Standards for medical identifiers, codes and messages needed to create an efficient computer-stored medical record. *J Am Med Inform Assoc* 1994;1(1):1-7.
13. Balaguer Perigüell E. Las nomenclaturas en Documentación clínica. Evolución de sus problemas. *Med Esp* 1974;71:191-200.
14. Feinstein AR. Unsolved scientific problems in the nosology of clinical medicine. *Arch Intern Med* 1988;148:2269-74.
15. Barona JL. La construcció sociocultural de les malalties. *Viure en Salut* 2000;(49):4-5.
16. Martin JS, Marsh JL. Current classification of fractures. Rationale and utility. *Radiol Clin North Am* 1997;35(3):491-506.

17. López Piñero JM, Terrada Ferrandis ML. *Introducción a la terminología médica.* Barcelona: Salvat, 1990.

18. Navarro FA. ¿Citocinas, citoquinas o citokinas? *Med Clin (Barc)* 2001;116(8):316-8.

19. Lancaster FW. *El control del vocabulario en la recuperación de información*. València: Universitat de València, 1995.

20. Chaumier J. *Analyse et langages documentaires*. París: Entreprise Moderne d'Édition, 1982.

21. Powsner SM, Wyatt JC, Wright P. Opportunities for and challenges of computerisation. *Lancet* 1998;352(9140):1617-22.

22. Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc* 1994;1(2):161-74.

23. Zweigenbaum P, Bouaud J, Bachimont B, Charlet J, Boisvieux JF. Evaluating a normalized conceptual representation produced from natural language patient discharge summaries. *J Am Med Inform Assoc* 1997;4(Suppl):590-4. Disponible en URL: http//www.biomath.jussieu.fr/~pz/Publications/biblio-pierre-pardate. [Citado 14 Sep 2001].

24. Brown SH, Miller RA, Camp HN, Guise DA, Walker K. Empirical derivation of an electronic clinically useful problem statement system. *Ann Intern Med* 1999;131:117-26.

25. Bolant Rodríguez A, Bosch Sánchez S, Gosálbez Pastor E, Marín Gómez M, Ortega Llavador JB, Sempere Soler J, Taberner Alberola F. *Manual de uso del codificador automático de urgencias hospitalarias.* Valencia: Generalitat Valenciana. Conselleria de Sanitat, 2000.

26. Bolant A, Puig-Moll J, Gosálbez E. Un sistema informático de ayuda a la codificación de diagnósticos. *Inform Salud* 1993;(5):165-9.

27. Gosálbez E, Bolant A, Puig-Moll J, Pérez-Salinas I. Sistema informatizado de codificación de procedimientos médicos. *Gestión Hosp* 1996;(2):19-26.

28. Grémy F. *Informatique médicale. Introduction a la méthodologie en médecine et santé publique.* París: Flammarion, 1987.

29. Lindberg DAB, Humphreys BL, McCray A. The Unified Medical Language System. *Meth Inf Med* 1993;32(4):281-91.

30. U.S. National Library of Medicine. National Institute of Health. Unified Medical Language System (UMLS) [páginaWeb]. Disponible en URL: http://www.nlm.nih.gov/research/umls. [citado 14 Sep 2001].

31. Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc* 1998;5(1):1-11.

32. McCray AT. The nature of lexical knowledge. *Meth Inf Med* 1998;37(4-5):353-60.

33. Kudla KM, Rallins MC. SNOMED: a controlled vocabulary for computer-based patient records. *J Am Health Inf Manag Assoc* 1998;69(5):40-4.

34. Rothwell DJ. SNOMED-based knowledge representation. *Meth Inf Med* 1995;34:209-13.

35. Lussier YA, Rothwell DJ, Côté RA. The SNOMED model: a knowledge source for the controlled terminology of the computerized patient record. *Meth Inf Med* 1998;37:161-4.

36. NHS Information Authority. *SNOMED Clinical Terms* [página Web]. Disponible en URL: http://www.coding.nhsia.nhs.uk/clin_term/snomedct/snomedct.asp. [citado 14 Sep 2001].

37. Booth N. What are the Read Codes? *Health Libr Rev* 1994;II:177-82.

38. Harding A, Stuart-Buttle C. The development and role of the Read Codes. *J Am Health Inf Manag Assoc* 1998;69(5):34-8.

39. GALEN Administrator, Medical Informatics Group, Department of Computer Science, University of Manchester. The Galen Project [página Web]. Disponible en URL: http://www.cs.man.ac.uk/mig/galen/index.html. [citado 14 Sep 2001].

40. Rector AL, Solomon WD, Nowlan WA, Rush TW, Zanstra PE, Claasen WM. A terminology server for medical language and medical information systems. *Meth Inf Med* 1995;34(1-2):147-57.