Revista Española de Documentación Científica
36(2), abril-junio 2013, e008
ISSN-L:0210-0614. doi: http://dx.doi.org/10.3989/redc.2013.2.915

ESTUDIOS / *RESEARCH STUDIES*

Identificación de indicios de descubrimientos científicos en artículos biomédicos mediante análisis de contenidos

Luciana Reis Malheiros*, Carlos Henrique Marcondes**

* Departamento de Fisiologia y Farmacología, Universidad Federal Fluminense, Niterói, Brasil. Correo-e: malheiro@vm.uff.br
** Departamento de Ciencia de la Información, Universidad Federal Fluminense, Niterói, Brasil. Correo-e: marcon@vm.uff.br

Recibido: 20-11-2011; 2ª version: 12-07-2012; Aceptado: 05-10 2012

Cómo citar este artículo/ Citation: Malheiros, L. R.; Marcondes, C. H. (2013). Identificación de indicios de descubrimientos científicos en artículos biomédicos mediante análisis de contenidos. *Revista Española de Documentación Científica*, 36(2):e008. doi: http://dx.doi.org/10.3989/redc.2013.2.915

Resumen: Este trabajo propone un método para la identificación de indicios de descubrimientos (ID) significativos que suponen avances relevantes en el conocimiento científico en el área biomédica. El método de identificación se realiza a través de la comparación de las conclusiones de artículos de esta área con el contenido de bancos de datos terminológicos públicos en la Web. Lo que se pretende reconocer ID presentes en un artículo antes incluso de que esté referenciado por la literatura, con el fin de prever las posibilidades de impacto que haya en el artículo. Se analizaron manualmente 89 artículos. Los resultados obtenidos indican si los contenidos de la conclusión de un artículo están pobremente representados en la ontología, esto puede que sea un indicio de descubrimiento significativo. Un indicio en favor de esta hipótesis es el hecho de que el artículo que marca el descubrimiento de la enzima telomerasa es de 1985, pero el término "telomerasa" sólo se incluyó en el MeSH tras 10 años.

Palabras clave: Representación del conocimiento; comunicación científica; descubrimiento científico; ontología.

Identification of evidence of scientific discoveries in biomedical articles through content analysis

Abstract: We report here a methodological proposal consisting of the comparison between the content of scientific articles, represented by the conclusion of the article in a format as phenomenon "1"- Relation – Phenomenon "2", with the content of a public Web-based ontology. This comparison was performed in order to identify traces of scientific discovery reported by the article even before its reference in the literature. Eighty-nine biomedical articles were manually analyzed. The results indicate that if the contents of the conclusion of an article are poorly represented in the ontology, this may be an indication of a significant discovery. One indication supporting this hypothesis is the fact that the article describing the discovery of the telomerase enzyme dates from 1985, but the term "telomerase" was only included in the MeSH ten years later, in 1995.

Keywords: Knowledge representation; scientific communication; scientific discover; ontology.

1. INTRODUCCIÓN

La publicación de artículos científicos en la Web es una actividad común en el medio científico y la mayoría de las revistas científicas posee una versión accesible en la Web. Sin embargo, los recursos de las tecnologías de la información (TIC) no suelen usarse directamente para procesar el conocimiento contenido en el texto de artículos científicos. Los artículos publicados en formato digital son "bases de conocimiento", pero, solamente, para la lectura humana. Existen dos barreras para el uso a gran escala de ese conocimiento: la cantidad de información disponible a través de la Web y el hecho de que el conocimiento está en un formato textual, de manera no estructurada, inadecuado para el procesamiento por programas de ordenador. Aún hoy, las revistas electrónicas están basadas en el modelo en papel.

Kuhn (2005) discute la importancia de las categorías conceptuales para la percepción de nuevos fenómenos, en el contexto de los cambios de paradigmas. Según el autor, establecer nuevas categorías y acuñar términos que las representen sería, por lo tanto, algunas de las características de los cambios de paradigmas científicos. Sin embargo, se debe considerar que un cambio de paradigma puede ocurrir sin la creación de nuevas categorías o fenómenos entre ellos. De esta manera, existirá siempre un hueco de tiempo entre la conceptualización de un nuevo descubrimiento y su representación como concepto en una terminologia.

¿Se pueden identificar indicios de nuevos descubrimientos (ND) a través del contenido de la literatura de determinado domínio científico? ¿Cómo?

Trabajamos hace años (Marcondes y otros, 2009) en la propuesta de un modelo de publicación de artículos científicos cuya intención es permitir que sus conclusiones sean "inteligibles" por programas de ordenadores. Artículos, no sólo publicados en el formato textual, sino que también tienen sus conclusiones identificadas, extraídas, formateadas como tripletes RDF (Research Description Framework), grabadas y publicadas en un formato procesable por máquina. Se puede decir que sería un subproducto del proceso de auto-publicación en el que los propios autores describirían sus conclusiones al someter el artículo a un sistema de publicación electrónica de una revista.

Nuestro enfoque respecto a la representación del conocimiento de las conclusiones de artículos científicos está basado en el hecho de que el conocimiento científico está constituido por aserciones hechas por los científicos en el texto de los artículos, expresando relaciones entre fenómenos o entre un fenómeno y sus características. Se consideran las relaciones como la unidad básica del conocimiento científico que sintetizan las conclusiones del artículo. A partir del momento en que se puedan extraer las conclusiones, marcadas como relaciones y grabadas en un formato procesable por máquina, será posible su procesamiento por agentes de software, proporcionando a los científicos nuevos medios de recuperar, comparar y evaluar dicho conocimiento.

En Marcondes (2011) se describe el prototipo de un sistema informático con una interfaz Web para el envío de artículos a revistas electrónicas que hace dos funciones: 1) a los autores les pide que, además de los metadatos convencionales, incluyan un texto corto con las conclusiones del articulo; el texto de las conclusiones es procesado por el sistema informático y formateado como tripletes (Research Description Framework) ; 2) cada concepto en los tripletes (Research Description Framework) es buscado por el término más próximo en UMLS (*Unified Medical Language System*); el resultado es mostrado al autor para que evalúe si el concepto de la conclusión está bien representado por el término encontrado en UMLS y si lo aprueba o no. La conclusión del artículo formateado en RDF y los resultados de la evaluación del autor son grabados para posterior procesamiento.

Una vez presentado en un formato que se pueda procesar por ordenador, las conclusiones de los artículos podrán ser comparadas por los programas con el conocimiento registrado en bancos de datos terminológicos públicos en la Web como el UMLS revelando inconsistencias, errores y quizás posibles indicios de descubrimientos. De esa manera es posible que un artículo científico, en el momento de su publicación en una revista electrónica y sin que, todavía, haya sido referenciado o citado, revele indicios que puedan indicar que en él se hace un descubrimiento importante.

Nuestra hipótesis es que existe una correlación entre un artículo cuya conclusión es representada de manera débil o representada solamente de modo genérico en bancos de datos terminológicos, como el UMLS, y el hecho de que esos artículos se refieran a descubrimientos científicos importantes.

2. OBJETIVOS

El objetivo de este trabajo es demostrar la viabilidad teórica y práctica de un método para identificar posibles descubrimientos importantes basándose en la comparación del contenido de las conclusiones de artículos científicos con contenidos terminológicos estandarizados registrados en ontologías o bancos de datos terminológicos públicos en la Web. Como objetivo secundario, se proponen métodos alternativos a los indicadores cientométricos de impacto basados en citas, métodos eficaces en el momento mismo de la publicación de un artículo, independientes de referencias *a posteriori* y que estén basados solamente en el contenido del artículo.

3. REFERENCIAL TEÓRICO

A fines del siglo XX, con el surgimiento de la *World Wide Web*, se ha vuelto cada vez más común

la publicación de artículos científicos en el formato digital. Las revistas científicas publicadas en la Web pueden ser una herramienta cognitiva cuyas potencialidades aún no se evaluaron totalmente. Aunque publicadas en la Web, las revistas electrónicas todavía están basadas en el modelo tradicional de las publicaciones en papel y no se utiliza todo el potencial del medio electrónico. El formato impreso es para que se lea, se evalue y se critique por personas; sin embargo requieren de un largo proceso de lectura, evaluación y citación por los pares para que los nuevos conocimientos, por fin, se incorporen al acervo de conocimiento público aceptado en un determinado campo.

En este proceso de comunicación del conocimiento científico hacer citaciones a otros artículos científicos no es sólo usual, sino también necesario. Hamilton (1990) relata, aún así, que el 55% de los artículos publicados en revistas indexadas por el ISI, de 1981 hasta 1985, no recibieron ninguna citación después de cinco años tras ser publicados. Y, además, los artículos que fueron citados, no lo fueron con mucha frecuencia; solamente el 42% de los artículos citados recibieron más de una citación.

De este modo, un artículo que demore en recibir citaciones puede formar parte de un grupo de artículos conocidos como de reconocimiento tardío (también llamado de descubrimiento prematuro o descubrimiento resistente), es decir, artículos que contribuyen de manera importante, pero que en un primer momento no recibieron la atención necesaria por parte de la comunidad científica. Con el paso del tiempo, el valor de un "artículo tardío" es (re)descubierto (Campanario, 1993).

A su vez, Niiniluoto (2007), critica severamente el uso de los indicadores cienciométricos como instrumentos para la detección del progreso científico, dice que "*they do not take into account the semantic content of scientific publications*".

De entre los factores que determinan que un artículo importante no recibiera la atención necesaria, se destacan: el artículo presentaría conclusiones que no corresponden a la teoría más aceptada por una determinada área; el autor del artículo es un investigador principiante y/o trabaja en una institución de investigación de poco prestigio; o además, el gran número de artículos publicados impediría que los artículos que traen nuevas ideas tuvieran relieve entre los que corroboraron el conocimiento ya establecido (Garfield, 1970).

El caso más exitoso de reconocimiento tardío es el artículo de Mendel sobre hibridación de plantas y publicado en 1865. El artículo fue citado pocas veces hasta ser "redescubierto" en 1900 (Garfield, 1970). Garfield ofrece el ejemplo de otros cinco artículos que se pueden considerar de reconocimiento tardío y que se identificaron a través del análisis de la frecuencia de citaciones. Él concluye su trabajo diciendo que el fenómeno de reconocimiento tardío parece ser poco usual.

Garfield retoma el tema y señala más artículos que estarían en esta categoría, alzando algunas cuestiones pertinentes como:

> *Is delayed recognition more prevalent among methods or concept papers? (...) Is there a difference over the past few decades, where de existence of improvement information retrieval methods has ostensibly made it more difficult to be unaware of relevant work? Or is there some fundamental delay factor that must inevitably affect the acceptance of new ideas via the educational-research process?* (Garfield, 1990)

En un último trabajo (Glänzel y Garfield, 2004) los autores reafirman que los casos de artículos que tienen reconocimiento tardío son pocos y que la mayoría de los artículos importantes están citados, como mucho, entre los primeros tres a cinco años de publicación. De los 60 artículos de reconocimiento tardío encontrados por ellos, el 43% eran del área de ciencia de la vida.

Fue Van Raan (2004) quien llamó "*Sleeping Beauties*" a los artículos con citas después de un largo período sin que les hagan referencias (son despertados por el príncipe'). Él estudió "las Bellas Durmientes" a partir de tres variables. La primera sería "la profundidad del sueño", medida por el número medio de citaciones recibidas en un determinado período de tiempo. Los artículos que recibieron, como máximo, una citación de media por año fueron considerados "en sueño profundo", los que recibieron entre una y dos citaciones de media por año se consideraron "en sueño leve". La segunda variable que se consideró fue el "tiempo del sueño", es decir la duración del período en el que los artículos recibieron como máximo, dos citaciones de media. Por último, se consideró la "intensidad del despertar"; o sea, el número medio de citaciones cuatro años después de "despertar".

Entonces, en un universo de cerca de un millón de artículos, él encontró 41 artículos que después de un "sueño profundo" de diez años recibieron, de media, seis ó siete citaciones en los cuatro años siguientes. Una crítica que el propio autor hizo a su trabajo es que había trabajado con varias áreas de conocimiento y que el patrón de citaciones de cada área es muy particular.

Además, la Web es un gran repositorio y distribuidor de informaciones sean de textos, imágenes o sonidos. A causa de la utilización de herramientas propias, cualquier persona puede encontrar esas informaciones con diferentes grados de dificultad, pues él/ella sabe reconocer su significado. El reto es hacer que los resultados y las conclusiones de investigación, como, por ejemplo, los encontrados en los artículos científicos, puedan ser "interpretados" permitiendo que los ordenadores puedan auxiliarnos en tareas más sofisticadas que demanden el procesamiento de dichos datos, disminuyendo la intervención humana y aumentando la precisión de las informaciones obtenidas. Par-

ticularmente, en el área biomédica, una enorme cantidad de información está disponible en formato digital como, por ejemplo, datos sobre la secuenciación genética (Stein, 2008), pero que todavía no están integrados en otras bases de datos, limitando su utilidad.

De esta forma, el objetivo de construcción de ontologías es el de registrar y almacenar conocimiento y permitir que múltiples sistemas y agentes "entiendan" el contenido de un recurso de la Web, y que puedan integrar este conocimiento con el contenido de otros recursos; el sistema o agente debe de ser capaz de interpretar la semántica de cada recurso (Jacob, 2003).

De Roure y otros (2001) enfatizan la importancia de la integración del conocimiento de diferentes fuentes, incluyendo artículos científicos publicados en la Web a los futuros ambientes de e-Science. Para lograr esta meta se necesita presentar el conocimiento en un formato procesable por máquina.

En esa dirección, uno de los esfuerzos de representación del conocimiento del área biomédica es el *Unified Medical Language System* (UMLS), un proyecto de la *National Library of Medicine* (NLM) que combina diversas fuentes terminológicas en un único instrumento. El UMLS posee una estructura jerárquica, o *Metathesaurus,* con cerca de 730.000 conceptos y más de 1 millón de nombres de conceptos. Está complementado por una estructura clasificatoria llamada *Semantic Network,* formada por clases de conceptos interrelacionados entre sí por tipos de relaciones.

Desde su creación existe una preocupación en añadir profesionales de áreas distintas para pensar sobre UMLS, así, bibliotecarios, científicos de la información, lingüistas, científicos de la computación, médicos, biomédicos, y otros, siempre han formado parte del equipo del UMLS (Humphreys y otros, 1998).

El objetivo del UMLS es el de "*facilitate the development of computer systems that behave as if they 'understand' the meaning of the language of biomedicine and health*" (National Library of Medicine, 2006, p.1). Para alcanzar este objetivo la NLM produce y distribuye bases de datos de la UMLS, nombradas UMLSKS (UMLS *Knowledge Sources*). Además de la UMLSKS, la NLM produce y distribuye también, softwares de apoyo que sirven de herramienta para que expertos en desarrollo de sistemas puedan crear o perfeccionar sistemas de información que procesen, creen, recuperen, integren y/o agreguen datos y/o informaciones biomédicas y de la salud.

No obstante, uno de los aspectos que más polémica generó en la construcción del UMLS fue la definición de cómo debería ser elaborado el *Metathesaurus*. No había acuerdo sobre la decisión de la NLM de construir el *Metathesaurus* a partir de la combinación de los conceptos de vocabularios fuentes. Sin embargo, la NLM argumentaba que no disponía de recursos para emprender la construcción de un vocabulario controlado tan extenso que pudiera atender a la demanda del UMLS (Humphreys y otros, 1998). La manera utilizada para la construcción del *Metathesaurus* implica que todos los conceptos, nombres y relaciones presentes en los diferentes vocabularios básicos, estén presentes en el *Metathesaurus ,* por ello

> *when two different source vocabularies use the same name for differing concepts, the Metathesaurus represents both of the meanings and indicates which meaning is present in which source vocabulary. When the same concept appears in different hierarchical contexts in different source vocabularies, the Metathesaurus includes all the hierarchies. When conflicting relationships between two concepts appear in different source vocabularies, both views are included in the Metathesaurus. […] the Metathesaurus does not represent a comprehensive NLM-authored ontology of biomedicine or a single consistent view of the world (except at the high level of the semantic types assigned to all its concepts*). (National Library of Medicine, 2008,)

Para Bondereider (2001) el Metathesaurus se puede considerar base de una ontología biomedica.

4. METODOLOGÍA

Artículos del área biomédica fueron elegidos como material empírico ya que suelen presentar una estructura más rígida, conteniendo: Introducción, Método, Resultados y Discusión (IMRD). Según Burrough-Boenisch (1999) "*scientists write in this forma not only to meet journals requirements but also to comply with the expectations of the scientific community*". También comenta que la mayoría de los manuales de redacción científica estimulan el uso de la estructura IMRD por considerarla la más adecuada para la organización del artículo científico. El ICMJE - *International Committee of Medical Journals Editors* (2008) dice que la estructura IMRD "*is not an arbitrary publication format but rather a direct reflection of the process of scientific discovery*".

En total, fueron analizados manualmente 89 artículos del área biomédica. Veinte artículos de la revista *Memórias del Intituto Owaldo Cruz* (MIOC), veinte del *Brazilian Journal of Medical and Biological Research* (BJMBR), veinte artículos que trataban de la investigación con terapia génica de células germinales y quince artículos de los ganadores del premio Lasker de 2006. El premio Lasker es otorgado anualmente y considerado tan importante como el Nobel, pese a ser menos conocido. De hecho, es considerado como uno de los premios que a veces precede al Nobel.

Los artículos del MIOC y del BJMBR fueron escogidos a través del portal Scielo utilizando la lista de artículos más visitados de cada uno de ellos.

Ambas revistas publican artículos en inglés y poseen un cuerpo editorial cualificado, con revisores nacionales e internacionales.

El primer grupo de artículos analizados se constituyó con artículos del MIOC que se edita desde 1909 y mantiene una excelente reputación nacional e internacional. Posteriormente, analizamos los artículos del BJMBR que se edita desde 1981 y que sustituyó a la *Revista Brasileira de Pesquisas Médicas e Biológicas*. Tanto el MIOC como el BJMBR están indexados por Scielo, LILACS, Medline e ISI/Thompson. En 2006, el factor de impacto para BJMBR fue de 1,075 y de 1,208 para el MIOC.

En la búsqueda de artículos que trajeran indicios de descubrimientos importantes, el tercer grupo de artículos analizados trataba de investigaciones sobre terapia génica de células germinales. La selección de los artículos de ese grupo se hizo a través de la lectura de tres artículos recientes de revisión del área (National Institutes of Health, 2006; Friel y otros, 2005; Bongso y Richards, 2004) en los que presentaban una visión histórica de la investigación de la terapia génica de células germinales, resaltando los avances más importantes, informaciones extremamente relevantes para esa investigación.

Buscando artículos que informaran de descubrimientos importantes, se eligió un último grupo de artículos. Ese grupo estaba compuesto de artículos que constaban en la bibliografía seleccionada de tres investigadores científicos - Elizabeth H. Blackburn, Carol W. Greider e Jack W. Szostak – ganadores, en 2006, del premio Albert Lasker de Investigación Médica Básica, que llevaron al descubrimiento de la telomerasa. Cada autor laureado proporcionó una lista de sus trabajos que creían más importantes y, de la unión de las tres listas, se obtuvieron los 15 artículos analizados. Proceden de varias revistas científicas como *Cell* y *Nature,* revistas con alto factor de impacto, 29,887 y 28,751, respectivamente.

El análisis del contenido de los artículos del premio se basó en los comentarios hechos por los propios autores sobre la mayoría de los artículos seleccionados. Esos comentarios forman parte de la revisión que los autores escribieron para la *Nature Medicine* (Blackburn y otros, 2006) con ocasión de la ceremonia del premio Lasker. En ella, los autores presentan la trayectoria de la investigación, resaltando los artículos que creían más importantes y especificando la contribución de cada uno de ellos. Estos artículos fueron denominados Grupo 1 de Telomerasa.

Catorce artículos adicionales sobre los desarrollos ulteriores de la investigación sobre la enzima telomerasa, de colaboradores próximos a sus descubridores, fueron seleccionados de la revisión hecha por Cech (2004) o de la linea de tiempo disponible en la base de datos Telomerasa. Estos artículos fueron denominados Grupo 2 de Telomerasa.

Debido a la estructura textual altamente formalizada de sus artículos, se seleccionaron revistas del área Biomédica. La mayor parte de los artículos del MIOC eran del área de la microbiología; los del BJMBR eran más heterogéneos predominando artículos de las áreas de fisiología y neurociencias; por fin, los artículos de terapia génica de células germinales y telomerasa que trataban de cuestiones relacionadas con la genética. Es importante enfatizar que la elección de esas revistas no se hizo en un único momento sino gradualmente, a lo largo del desarrollo de la investigación, buscando siempre artículos que trajeran descubrimientos científicos importantes, objeto de ese trabajo.

El proceso de análisis de los artículos se hizo en dos etapas. En un primer momento, el grupo debería intentar identificar en el texto cual era la principal conclusión presentada por los autores. Para esta tarea se echó mano, también, de artículos de revisión que hacían referencia al trabajo analizado. Identificada la principal conclusión, se discutía la mejor manera de expresarla sintéticamente en la forma de antecedente (un concepto que se refiere a un fenómeno), una relación semántica y una consecuencia (otro concepto que se refiere a un fenómeno o una característica del fenómeno expreso en el antecedente). Como, por ejemplo, el análisis del artículo "*A mutant with a defect in telomere elongation leads to senescence in yeast*" (Lundblad y Szostak, 1989) lleva a la conclusión sintetizada en la siguiente afirmación: El acortamiento del telómero causa senescencia celular. O esquemáticamente:

Antecedente: acortamiento del telómero

Relación: causa

Consecuencia: senescencia celular

Los descriptores MeSH de ese artículo son: aging/physiology*, alleles, amino acid sequence, base sequence, cell survival, chromosome aberrations, chromosome disorders, chromosome/physiology*, cloning/molecular, DNA/analysis, molecular sequence data, mutation*, phenotype, Saccharomyces cerevisiae/genetics*

El artículo al ser publicado, es indexado casi inmediatamente. Considerando que la indexación fue hecha con los mejores términos disponibles en la época de publicación, y establecidos el Antecedente, la Relación y la Consecuencia, se verificó en qué grado estos elementos estaban representados en la indexación MeSH del artículo. A eso, damos el nombre de mapeado. Si todos los elementos (el antecedente, relación y consecuencia) fueron mapeados en el UMLS, el artículo era considerado completamente mapeado: CM. Si el antecedente o la consecuencia fueron mapeados, el artículo era considerado parcialmente mapeado – PM. Si ninguno de los elementos fueron mapeados, el artículo era considerado no mapeado – NM.

Cada artículo analizado fue también clasificado bajo cuatro clases distintas según el tipo de razonamiento desplegado en él y la existencia de una hipótesis en el texto que lo orientase; se encontraron las siguientes clases de razonamiento:

- EI - artículos experimentales hipotéticos inductivos, los que a partir de una hipótesis original y dos resultados de un experimento sacan o inducen nuevas conclusiones;
- ED - artículos experimentales hipotéticos deductivos, los que a partir de una hipótesis planteada por otro autor, desarrollan experimentos que solo confirman la hipótesis del otro autor;
- EE - artículos experimentales exploratorios, no parten de ningúna hipótesis previa y simplemente describen o caracterizan un nuevo fenómeno;
- TA - artículos teóricos abductivos, no desarrollan ningún experimento, sino solo proponen teóricamente una nueva hipótesis.

La descripción detallada de estas clases de razonamiento se encuentra en Marcondes y otros (2009) y Marcondes (2011). Las clases de razonamiento como TA y EE están directamente ligadas al foco de esta investigación, la identificación de rasgos de descubrimientos científicos. Aunque "novedad científica" sea una noción sin una definición exacta y la literatura muestre intentos de definición más cualificada, para efectos de esta investigación, vamos a considerar indicios de novedad científica el mapeado parcial o el no mapeado de conceptos de la conclusión del artículo en ontologías de acceso público, en el mismo dominio científico del artículo.

5. RESULTADOS

De los 89 artículos analizados, los grupos que reportaron novedades científicas – los artículos que relatan el descubrimiento y los desarrollos ulteriores de la investigación sobre la enzima telomerasa (TELOMERASA 1 y 2) – seguidos por los artículos sobre células madre, han obtenido las menores tasas de no mapeo (NM) –, como se muestra en la Tabla I: artículos que han obtenido no mapeo son 45% de estos grupos, el grupo de artículos sobre células madre obtuvo 20%, el grupo del BJMBR obtuvo el 30% y el grupo del MIOC el 0% de no mapeos. Este mismo grupo de artículos muestra los pasos hasta el descubrimiento de la enzima telomerasa, un descubrimiento científico importante. El Premio Lasker de Medicina es considerado una anticipación del Premio Nobel: de hecho, los ganadores de Premio Lasker de Medicina de 2006 son también los ganadores del Nobel de 2009. Dentro de este grupo de artículos ninguno (0%) ha obtenido completo mapeo (CM) de todos los elementos de su conclusión (Antecedente o Consecuencia) en términos del *Medical Subject Headings* (MeSH); 6 de 15 (40%) han obtenido mapeo parcial (MP) y 9 de 15 (60%) no han obtenido ningún mapeo (NM).

En el grupo de artículos sobre células madre ningún artículo obtuvo completo mapeo (CM) en los elementos de su conclusión; 16 de 20 (80%) obtuvieron mapeo parcial (PM) y 4 de 20 (20%) no obtuvieron ningún mapeo (NM). En los artículos del grupo BJMBR, 14 de 20 (70%) han obtenido mapeo completo (CM) o parcial (PM) y 6 (30%) han obtenido no mapeo (NM). Los artículos del grupo MIOC obtuvieron el más grande grado de mapeo (CM y PM); ningún artículo (0%) obtuvo no mapeo (NM).

Tabla I. Resultados del mapeo de conceptos de las conclusiones en términos MeSH, por grupo de artículos

Artículos analizados	MIOC	BJMBR	CELULAS MADRE	TELOMERA-SA 1	TELOMERA-SA 2	TELOMERASA 1 + 2	TOTAL
CM-Completamente mapeados	7 (35%)	3 (15%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	10
PM-Parcialmente mapeados	13 (65%)	11 (55%)	16 (80%)	6 (40%)	10 (71%)	16 (55%)	54
NM-No mapeos	0 (0%)	6 (30%)	4 (20%)	9 (60%)	4 (29%)	13 (45%)	25
Total de artículos	20	20	20	15	14	29	89

Rev. Esp. Doc. Cient., 36(2), abril-junio 2013, e008. ISSN-L: 0210-0614. doi: http://dx.doi.org/10.3989/redc.2013.2.915

La Tabla II muestra el grupo de artículos de las *key publications* del Premio Lasker de Medicina de 2006 –Telomerasa 1– ordenados cronológicamente, cada artículo con su grado de mapeo y su tipo. El orden cronológico permite observar que artículos anteriores a la publicación del artículo que marca el descubrimiento de la enzima telomerasa y su nombramiento en 1985 son todos del tipo experimental exploratorio (EE) y carecen de mapeo de sus conclusiones en términos MeSH (NM). Los artículos de tipo experimental exploratorio parecen caracterizar los primeros pasos hasta el descubrimiento de un nuevo fenómeno. Después del descubrimiento y nombramiento de la telomerasa en 1985 aparecen los primeros artículos de tipo experimental inductivo y experimental deductivo. Después de 1986 también aparecen los primeros artículos de mapeo parcial (PM).

Tabla II. Grupo 1 de Telomerasa, artículos ordenados cronológicamente con grado de mapeo y de clase de razonamiento

Año	Titulo	Mapeo	Clase/razón
1978	Blackburn, E. H. & Gall, J. G. A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. *J. Mol. Biol*. 1978, 120:33-53.	NM	EE
1982	Szostak, J. W. & Blackburn, E. H. Cloning yeast telomeres on linear plasmid vectors. *Cell* 1982, 29:245-255.	NM	EE
1983	Murray, A. W. & Szostak, J. W. Construction of artificial chromosomes in yeast. *Nature* 1983, 305:189-193.	NM	EE
1984 JAN	Shampay, J., Szostak, J. W., Blackburn, E. H. DNA sequences of telomeres maintained in yeast. *Nature* 1984, 310:154-157.	NM	EE
1984 MAY	Dunn, B. L., Szauter, P., Pardue, M. L., Szostak, J. W. Transfer of telomere-adjacent sequences to linear plasmids by recombination. *Cell* 1984, 39:191-201.	NM	EE
1985[1]	Greider, C. W., & Blackburn, E. H. Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. *Cell* 1985, 43:405-413.	PM	EE
1987	Greider, C. W. & Blackburn, E. H. The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. *Cell* 1987, 51:887-898.	NM	EE
1989 JAN	Lundblad V. & Szostak, J. W. A mutant with a defect in telomere maintenance leads to senescence in yeast. *Cell* 1989, 57:633-643.	PM	EI
1989 NOV	Greider, C. W., & Blackburn, E. H A telomeric sequence in the RNA of Tetrahumena telomerase required for telomere repeat synthesis. *Nature* 1989, 337:331-337.	NM	EE
1990	Yu, G. L, Bradley, J. D., Attardi, L.D. and Blackburn, E. H. In vivo alteration of telomere sequences and senescence caused by mutated Tetrahymena telomerase RNAs. *Nature* 1990, 344:126-132.	PM	ED
1992	Allsopp, R. C., Vaziri, H., Patterson, C., Goldstein, S., Younglai, E.V., Futcher, C. W., Greider, C. W., Harley, C. B. Telomere length predicts the replicative capacity of human fibroblasts. *Proc. Natl. Acad. Sci*. USA 1992, 89:10114-10118.	PM	ED
1993	Prowse, K. R., Avilion, A. A., Greider, C. W. Identification of a nonprocessive telomerase activity from mouse cells. *Proc. Natl. Acad. Sci*. USA 1993, 90:1493-1497.	PM	ED
1995[2]	McEachern, M. J. & Blackburn, E. H. Runaway telomere elongation cause by telomerase RNA mutations. *Nature* 1995, 376:403-409.	PM	EI
1999	Rudolph, K. L., Chang, S, Lee, H.W., Blasco, M., Gottlieb, G., Greider, C. W., and DePinho, R. A. Longevity, stress response, and cancer in aging telomerase deficient mice. *Cell* 1999, 96:701-716	PM	ED
2001	Kim, M. M., Rivera, M. A., Botchkina, I. L, Shalaby, R., Thor, A. D., Blackburn, E. H. A low threshold level of expression of mutant-template telomerase RNA is sufficient to inhibit tumor cell growth. *Proc. Natl. Acad. Sci*. USA 2001, 98:7982-7987	NM	ED

1 Este artículo marca el descubrimiento de la enzima telomerasa y la fijación de su nombre científico.
2 1995 marca la entrada del término telomerasa en MeSH, ver en http://www.nlm.nih.gov/cgi/mesh/2011/MB_cgi?mode=&term=Telomerase&field=entry.

La Tabla III es una continuación de la Tabla II: los artículos relatan descubrimientos sucesivos y complementarios al descubrimiento de la telomerasa y de su entrada como concepto en MeSH. 9 de un total de 14 son artículos experimentales exploratorios, mientras que 5 de un total de 14, son de clase de razonamiento experimental inductivo, una clase que va más delante de la simple caracterización-descripción de un fenómeno, proponiendo una (nueva) relación entre dos fenómenos, un numero más grande que en el anterior grupo.

La entrada de la telomerasa como concepto en MeSH ocurre solamente en el año 1995. De un total de 29 artículos de ambos grupos, 12 han sido publicados antes de 1995 y 17 después. De los artículos publicados antes, 7 de 12 (58%) han obtenido no mapeo (NM); de los 17 artículos publicados después, solamente 5 de 17 (29%) han obtenido no mapeo (NM), un decrecimiento porcentual de 29. En el grupo de 12 artículos publicados antes del año de 1996, 8 de 12 (67%) son experimentales exploratorios y los restantes 4 de 12 (33%) son experimentales inductivos o experimentales deductivos, un decrecimiento porcentual de 34.

Las Tablas II y III permiten comparar cronológicamente el mapeo con la clase de razonamiento

Tabla III. Grupo 2 de Telomerasa, artículos ordenados cronológicamente con grado de mapeo y de clase de razonamiento

Año	Titulo	Mapeo	Clase/razón
1998	Bodnar, A.G., Ouellette, M., Frolkis, M., Holt, S.E., Chiu, C.P., Morin,G.B., Harley, C.B., Shay, J.W., Lichtsteiner, S., and Wright, W.E. Extension of life-span by introduction of telomerase into normal human cells. *Science* 1998, 279: 349–352.	PM	EI
1999	Mitchell, J.R., Wood, E. & Collins, K. A telomerase component is defective in the human disease dyskeratosis congenita. *Nature* 1999, 402: 551–555.	PM	EI
2000	Chen, J.-L., Blasco, M.A., and Greider, C.W. Secondary structure of vertebrate telomerase RNA. *Cell* 2000, 100: 503–514.	NM	EE
2000	Tzfati, Y., Fulton, T.B., Roy, J., and Blackburn, E.H. Template boundary in a yeast telomerase specified by RNA structure. *Science* 2000, 288:863–867.	NM	EE
2001	Vulliamy, T., Marrone, A., Goldman, F., Dearlove, A., Bessler, M., Mason, P.J., and Dokal, I. The RNA component of telomerase is mutated in autosomal dominant dyskeratosis congenita. *Nature* 2001, 413:432–435.	PM	EI
2001	Hermann, M. T., Strong, M. A. Hao, L. Y., Greider, C. W. (2001). The shortest telomere, not average telomere length, is critical for cell viability and chromosome stability. *Cell*, 107(1), 67-77.	PM	EE
2002	Chen, J.-L., Opperman, K.K., and Greider, C.W. A critical stem-loop structure in the CR4–CR5 domain of mammalian telomerase RNA. *Nucleic Acids Res*. 2002, 30:592–597.	PM	EE
2002	Seto, A.G., Livengood, A.J., Tzfati, Y., Blackburn, E.H., and Cech, T.R. (2002). A bulged stem tethers Est1p to telomerase RNA in budding yeast. *Genes Dev*., 2800–2812.	NM	EE
2003	Ly, H., Xu, L., Rivera, M.A., Parslow, T.G., and Blackburn, E.H. A role for a novel "trans-pseudoknot" RNA-RNA-interaction in the functional dimerization of human telomerase. *Genes Dev*. 2003, 17: 1078–1083.	PM	EE
2003	Loayza, D., and de Lange, T. POT1 as a terminal transducer of TRF1 telomere length control. *Nature* 2003, 423:1013–1018.	NM	EE
2003	Chen, J.-L., and Greider, C.W. Template boundary definition of hTERT. *Genes Dev*. 2003, 17:2747–2752.	PM	EE
2005	Armanios, M. et al. Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita. *Proc. Natl. Acad. Sci*. USA 2005, 102:15960–15964.	PM	EE
2005	Hao, L.Y. et al. Short telomeres, even in the presence of telomerase, limit tissue renewal capacity. *Cell* 2005, 123: 1121–1131.	PM	EI
2007	Armanios, M. Y, Chen, J. J., Cogan, J. D., Alder, J. K., Ingersoll, R. G., Markin, C., Lawson, W. E., Xie, M, Vulto, I, Phillips, J. A., Lansdorp, P. M., Greider, C. W., Loyd, J. E. Telomerase mutations in families with idiopathic pulmonary fibrosis. *N Engl J Med*. 2007, 356(13):1317-26.	PM	EI

encontrado en cada articulo. Los resultados presentados tienen las siguientes características: primero se empieza a caracterizar en artículos experimentales exploratorios hasta ser apropiado por completo; posteriormente se relaciona a través de causa-efecto con otros fenómenos en artículos de género experimental inductivo. El cambio cualitativo se realiza por la completa caracterización del nuevo fenómeno y su encuadramiento en el marco conceptual de una disciplina científica. En el caso del descubrimiento de la telomerasa esto se da por la identificación de la telomerasa como *una enzima* y se la nombra en el artículo de Greider y Blackburn de 1985.

Se necesitaría estudiar otros casos similares y bien documentados por sus autores como el descubrimiento de la telomerasa para verificar si el mismo estándar se repite. El método propuesto nos remite a Kuhn (2005) cuando afirma que lel descubrimiento implica un proceso de asimilación conceptual amplio. También es coherente con la afirmación de Kuhn (2005) sobre los periodos pre -paradigmáticos en los que falta una terminología precisa y, además, consensual. El autor aún reitera que, desde lo cognitivo, son necesarios nuevos conceptos para manejar nuevos paradigmas; un nuevo paradigma necesita un completo sistema conceptual (y terminológico) para describirlo.

6. CONCLUSIONES

El objetivo de este trabajo es demostrar la viabilidad de un método que permite comparar las conclusiones de artículos científicos con el conocimiento registrado en ontologías o bancos de datos terminológicos públicos en la Web a fin de identificar posibles descubrimientos importantes. En el momento, no disponemos de una ontología altamente formal en el área biomédica, pero, se cree que con el desarrollo de esta área, el método aquí propuesto podrá apuntar indicios de posibles descubrimientos científicos de manera más precisa. Se usó aquí el MeSH como una herramienta, a falta, de momento, de otra mejor.

Los resultados indican que el grado de éxito/no éxito para el mapeo de la representación de la conclusión MeSH tiene correlación con el hecho de que los artículos relaten descubrimientos científicos importantes. Parece metodológicamente posible proponer un procedimiento en el que los autores expresen su principal conclusión de manera sintética y que sea automáticamente procesada y comparada con el conocimiento científico ya previamente establecido y representado en ontologías o bancos de datos terminológicos públicos.

La creciente cantidad de artículos que se publican constantemente, en especial en el área biomédica, vuelve mucho más difícil y torpe el proceso de identificación por investigadores de posibles artículos relevantes, su lectura, evaluación, crítica y eventual citación.

Sería muy importante el desarrollo de indicadores que tomasen en cuenta el contenido de un artículo científico y pudieran hacerlo directamente en el momento de la publicación del artículo, sin tener que esperar un largo tiempo hasta que el artículo sea citado. Tales indicadores, asociados a un método automático de envío de artículos, como el propuesto en Marcondes (2011), que pueda también apuntar indicios de novedad, permitiría optimizar este proceso, para que la atención del investigador o del gestor de Ciencia y Tecnología pueda encontrarse con artículos que sean potencialmente relevantes.

Se debe considerar que la indexación de los artículos no se hace por los autores, que conocen mejor lo que se está relatando y la contribución que están haciendo a la ciencia. La indexación se hace posteriormente a la publicación, cuando los artículos son incluidos en bases de datos o repositorios como el Medline o PubMed.

De esta manera, un nuevo descubrimiento científico puede crear nuevos conceptos para los cuales los correspondientes términos aún no hayan sido incluidos en bases de datos terminológicas como el UMLS. Así, existe un retraso entre el descubrimiento de un fenómeno, o concepto y la actualización del UMLS. Eso se percibe fácilmente cuando se compara el desfase entre las palabras claves del autor en artículos biomédicos con los descriptores del MeSH atribuidos al artículo cuando éste es depositado en bibliotecas digitales como el PubMed. Un indicio a favor de esta hipótesis es el hecho de que, entre los artículos analizados del grupo que informa del descubrimiento de la enzima telomerasa - ganadores del premio Lasker de Medicina del año de 2006 y del premio Nobel de Medicina del año de 2009 - el artículo que marca el descubrimiento de la enzima es de 1985 (Greider y Blackburn, 1985), pero el término telomerasa sólo se incluyó en el MeSH tras 10 años.

Se cree que hay un gran potencial investigativo en lo investigado, y un retraso entre el descubrimiento de un nuevo fenómeno y su integración en su sistema conceptual y terminológico de un dominio científico. Similar a lo ocurrido en el descubrimiento de la telomerasa, en 1981 un informe del Center of Disease Control and Prevention, USA (CDC, 1981) recogió cinco casos de pneumocystis carinii pneumonia (PCP) entre hombres jóvenes de Los Angeles, uno de los primeros relatos de la enfermedad que se conocería como el SIDA. De acuerdo con la National Library of Medicine de EUA, un término para el SIDA no entró en el MESH hasta 1983. Cambios científicos implican nuevos sistemas conceptuales y hay un retraso temporal entre un nuevo descubrimiento y su representación en terminologías científicas.

Hace falta poner de relieve que, en algunos casos, la "novedad científica" no está acompañada de la creación de nuevos términos sino, por ejemplo, por la manera como dos fenómenos se relacionan.

Se conjetura que con el crecimiento de las ontologías como nuevos artefactos científicos (Smith, 2008), probablemente habrá nuevos procesos de validación /ratificación científicos. Además, las ontologías también están evolucionando para una mayor formalización y necesitarán de nuevos métodos de actualización y mantenimiento (Williams y Anderson, 2003).

Lo mismo se puede decir de los artículos científicos publicados en formato digital: tan pronto sean publicados en un formato más completo y formal, esto posibilitará el procesamiento de una conclusión y comparación con ontologías públicas de la Web, conforme aquí se propone.

Se cree que el método propuesto, después de totalmente automatizado e implementado, pueda convertirse en una herramienta más de evaluación de la producción científica y complementar a los ya tradicionales métodos bibliométricos y cienciométricos.

7. AGRADECIMIENTOS

Al Profesor Antonio Hernández-Pérez, de la Universidad Carlos III de Madrid, por sus valiosos comentarios y sugerencias.

Este artículo esta basado en los datos de la tesis de doctorado de Malheiros (2010).

8. BIBLIOGRAFÍA

Blackburn, E. H.; Greider, C. W.; Szostak, J. W. (2006). Telomeres and telomerase: the path from maize, *Tetrahymena* and yeast to human cancer and aging. *Nature Medicine*, vol.12 (10), VII-XII.

Bondereider, O. (2001). Medical ontology research. *Report to the board of scientific counselors of the Lister Hill National Center for Biomedical Communications*. Disponible en <http://mor.nlm.nih.gov/> [consultado el 06 de enero de 2008].

Bongso, A., y Richards, M. (2004). History and perspective of stem cell research. *Best Practice & Research Clinical Obstetrics & Gynaecology,* vol.18 (6), 827-842.

Burrough-Boenish, J. (1999). International reading strategies for IMRD articles. *Written Communication* , vol.16 (3), 296-316.

Campanario, J. M. (1993). Consolation for scientists: sometimes it is hard to publish papers that are later highly-cited. *Social Studies of Science*, vol. 23, 342-362.

CDC (1981). Pneumocystis pneumonia. Los Angeles. *MMWR*, vol. 30 (21), 1-3.

Cech, T. R. (2004). Beginning to understand the end of chromossomes. *Cell*, vol. 116, 273-279.

De Roure, D.; Jennings, N.; Shadbolt, N. (2001). Research agenda for the Semantic Grid: a future s-Science infrastructure. *Report Commissioned for EPSRC/DTI Core e-Science Programme*. p.78.

Friel, R.; Sar, S.; Mee P. (2005). Embryonic stem cells: understanding their history, cell biology and signalling. *Advanced Drug Delivery Reviews*, vol. 57 (13), 1894-1903.

Garfield, E. (1970). Would Mendel´s work have been ignored if the Science Citation Index® was available 100 years ago? *Essays of an Information Scientist*, vol. 1, 69-70.

Garfield, E. (1990). More delayed recognition. Part 2. From inhibin to Scanning Electron Microscopy. *Essays of an Information Scientist* ,13, 68-74.

Glänzel, W.; Garfield, E. (2004). The myth of delayed recognition. *The Scientist*, vol. 18 (11), 8.

Greider, C. W.; Blackburn, E. H. (1985). Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. *Cell*, vol. 43, 405-413.

Hamilton, D.P. (1990). Publishing by – and for? – the numbers. *Science*, vol. 250 (4986), 1331-32.

Humphreys, B. L.; Lindberg, D. A. B.; Schoolman, H.M.; Barnett, G. O. (1998). The Unified Medical Language System: an informatics research collaboration. *Journal of the American Medical Informatics Association*, vol. 5 (1), 1-11.

International Committee of Medical Journal Editor (ICMJE). (2008). Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication. *ICMJE*, 1-16. Disponible en <http://www.icmje.org/manuscript_1prepare.html> [consultado el 22 de noviembre de 2011].

Jacob, E. K. (2003). Ontologies and the Semantic Web. *Bulletin of the American Society for Information Science and Technology*, vol. 29 (4), 19-22.

Kuhn, T. S. (2005). *A estrutura das revoluções científicas*. (9ª ed.). São Paulo: Perspectiva. p. 260.

Lundblad, V.; Szostak, J.W. (1989). A mutant with a defect in telomere elongation leads to senescence in yeast. *Cell*, vol. 57 (4), 633-643.

Malheiros, L. R. (2010). *A identificação de traços de descobertas científicas pela comparação do conteúdo de artigos em Ciências Biomédicas com uma ontologia pública*. Tese (Doutorado em Ciência da Informação). Programa de Pós-Graduação em Ciência da Informação convênio UFF/IBICT, Niterói.

Marcondes, C. H. A semantic model for scholarly electronic publishing. (2011). Proceedings of the 1st International Workshop on Semantic Publication. Hersonissos, Crete: Greece, 721. ISSN: 1613-0073. Disponible en <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-721/.> [consultado el 30 de maio de 2011].

Marcondes, C. H.; Mendonça, M.A.R.; Malheiros, L.R.; Costa L.C.; Santos T.C.P. (2009). Ontological and conceptual bases for a scientific knowledge model in biomedical articles. *RECIIS*, vol. 3(1), 19-30, 2009. Disponible en <http://www.reciis.cict.fiocruz.br/index.php/reciis/article/view/240/251.> [consultado el 8 de abril de 2009].

National Institutes of Health. (2006). *The Human Embryonic Stem Cell and the Human Embryonic Germ*

Cell. Disponible en http://stemcells.nih.gov/. [consultado el 8 de marzo de 2006].

National Library of Medicine (2006). *Unified Medical Language System – Fact sheet.* Disponible en <http://www.nlm.nih.gov/pubs/factsheets/umls.html.> [consultaddo el 15 de noviembre de 2011].

National Library of Medicine. (2008). *Unified Medical Language System - Metathesaurus*. Disponible en <http://www.nlm.nih.gov/research/umls/meta2.html.> [consultado el 4 de enero de 2008].

Niiniluoto, I. (2007). Scientific Progress. En: Zalta, E.N. (editor). *The Stanford Encyclopedia of Philosophy*. Disponible en <http://plato.stanford.edu/archives/fall2008/entries/scientific-progress/.> [consultado el 1 de febrero de 2008].

Smith, B. Ontology (Science). (2008). *Nature Precedings*. Disponible en <http://hdl.handle.net/10101/npre.2008.2027.2.> [consultado el 1 de agosto de 2009].

Stein, L. D. (2008). Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nature Reviews Genetic*, vol. 9, 678-688.

Van Raan, A. F. J. (2004). Sleeping Beauties in Science. *Scientometrics*, vol. 59 (3), 467-472.

Williams, J.; Anderson W. (2003). Bringing ontology to the Gene Ontology. *Comparative and Functional Genomics*, vol. 4, 90–93. Disponible en <http://hindawi.com/GetPDF.aspx?doi=10.1002/cfg.253.> [consultado el 31de julio de 2009].