EDICIC
Asociación de Educación e Investigación en Ciencia de la Información de Iberoamérica y el Caribe

IDENTIFICACIÓN DE LOS RASGOS DE DESCUBIERTAS CIENTÍFICAS EN ARTÍCULOS BIOMÉDICOS

Luciana Reis Malheiros[1] e Carlos Henrique Marcondes[2]

[1]Profesora Adjunta del Depto. de Fisiología y Farmacología – Universidade Federal Fluminense (UFF) – Brasil

[2]Profesor Asociado del Depto. de Ciencia de la Información – Profesor del Programa de Posgrado en Ciencia de la Información – UFF – Brasil

RESUMEN

Este trabajo propone un método para la identificación de indicios de descubiertas significativas (ID) en el área biomédica a través de la comparación de la principal conclusión de artículos de esta área con el contenido de una ontología pública en la Web. De esta manera, se hace posible reconocer ID relatado en el artículo aunque antes de que se lo haga referencia por la literatura. Fueron analizados manualmente 75 artículos. Los resultados obtenidos indican que si los contenidos de la conclusión de un artículo están pobremente representados en la ontología, esto puede ser un indicio de descubierta significativa. Un indicio a favor de esa hipótesis es el hecho de que el artículo que marca la descubierta de la enzima telomerase es de 1985, pero el término “telomerase” sólo se incluyó en el MeSH tras 10 años.

Palabras-Clave: Representación del Conocimiento; Comunicación Científica; Descubierta Científica; Ontología.

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 Web public ontology. This comparison was performed in order to identify traces of scientific discovery reported by the article even before the reference in the literature. Seventy-five 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 is from 1985, but the term “telomerae” was included in the MeSH only in 1985, ten years later.

Keywords: Knowledge Representation; Scientific Communication; Scientific Discovery; Ontology.

1 INTRODUCCIÓN

La publicación de artículos científicos en la Web es una actividad común en el medio científico y la mayoría de los periódicos científicos posee una versión accesible en la Web. Sin embargo, los recursos de la tecnología de la información (TI) no se suelen usarse directamente para procesar el conocimiento continuo en texto de artículos científicos. Artículos publicados electrónicamente son “bases del conocimiento”, pero, solamente, para la lectura humana. Existen dos barreras para su uso en larga 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 computador. Aún hoy, los periódicos electrónicos están basados en el modelo del periódico en papel.

Kuhn (2005, p.149) discute la importancia de las categorías para la percepción de nuevos fenómenos, en el contexto de los cambios de paradigmas y dice: “Aunque, después que la experiencia en curso haya fornecido las categorías adicionales indispensables, se fue capaz de percibir las cartas anómalas […]” Establecer nuevas categorías y acuñar términos que las representen sería, por lo tanto, algunas de las características de los cambios de los paradigmas científicos. Sin embargo, se debe considerar que un cambio de paradigma puede ocurrir sin la ilusión de nuevas categorías o fenómenos entre ellos. De esta manera, existirá siempre un hueco de tiempo entre la conceptuación de una nueva descubierta y su representación como concepto en una terminología.

¿De qué manera indicios de descubiertas importantes (ID) pueden ser identificadas?

¿Artículos que traigan ID tendrán el contenido de sus conclusiones bien representado en ontologías públicas del mismo dominio de conocimiento del artículo?

¿Nuevos conceptos o fenómenos recién acuñados serán inmediatamente representados en esas ontologías?

Se cree que se puede hacer un avance en el área de publicación electrónica. Trabajamos hace años (MARCONDES *et al*., 2009) en la propuesta de un modelo de publicación de artículos científicos cuya propuesta es permitir que sus conclusiones

sean “inteligibles” por programas de ordenadores. Artículos, no sólo son publicados en el formato textual, sino que también tienen sus conclusiones identificadas, extraídas, grabadas y publicadas como instancias de una ontología en un formato procesable por máquina. Lo que se puede decir ser 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 un periódico.

Nuestro abordaje 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 consideró las relaciones como la unidad básica del conocimiento científico y que sintetizarían las conclusiones del artículo. A partir del momento en que se pueda extraer las conclusiones, marcadas como relaciones y grabadas en un formato procesable por máquina, será posible su procesamiento por agentes de software, forneciendo a los científicos nuevos medios de recuperar, comprar y evaluar dicho conocimiento.

Una vez presentada en un formato que se pueda procesar en máquina las conclusiones de los artículos podrán ser comparadas por los programas con el conocimiento registrado en ontologías públicas en la Web relevando, entonces, inconsistencias, errores y posibles indicios de descubiertas. De esa manera es posible que un artículo científico, en el momento de su publicación en un periódico electrónico y sin que, todavía, haya sido referenciado o citado, releve indicios que puedan indicar que en él se hace una descubierta 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 (o *Unified Medical Language System*), y el hecho de que esos artículos se reportan a descubiertas científicas importantes.

Eso se percibe fácilmente cuando se compara el desfase entre las palabras claves del autor en artículos biomédicos con los descriptores del *Medical Subject Headings* (MeSH) atribuidos al artículo cuando este es depositado en bibliotecas digitales como el PubMed. Un indicio a favor de esa hipótesis es el hecho de que, entre los artículos analizados del grupo que reporta la descubierta de la enzima

telomerase, el artículo que marca la descubierta de la enzima es de 1985 (GREIDER; BLACKBURN, 1985), pero el término telomerase sólo se incluyó en el MeSH tras 10 años.

El objetivo de este trabajo es demostrar la viabilidad de un método que compare las conclusiones de artículos científicos con el conocimiento expreso en ontologías públicas en la Web para identificar posibles descubiertas importantes.

2 REFERENCAL TEÓRICO

La comunicación científica formal es consecuentemente, la creación de los periódicos científicos que se relaciona directamente a la creación de las sociedades científicas. La *Royal Society,* por ejemplo empezó a editar en 1665 las *Philosophical Transactions,* periódico hasta hoy se edita. En la introducción del primer número Oldenburg (1665, p.1, bastardilla nuestra) dice: “Nada importa más para que se promueva el desarrollo de las Cuestiones Filosóficas que **comunicar** […] los estudios y los descubrimientos más recién en el área”. Desde su inicio, había un espacio en el periódico para que fueran publicadas cartas de investigadores, comentando algún artículo que había sido publicado.

Con la introducción del periódico, tuvo inicio el proceso de formalización de la comunicación científica y el registro de las investigaciones se hizo disponibles por largo período de tiempo para un público mucho más amplio que aquél que discuta las cuestiones científicas en cartas personales (MEADOWS, 1999).

Para Ziman (1979, p.118) la creación de la revista científica “[…] tuvo una importancia mayor que cualquier otra iniciativa de las Sociedades Reales y Academias Nacionales […]”. El periódico es una publicación que, como revela el nombre, debe ser regular y además posibilitar la publicación de un gran número de investigaciones. De nuevo, Ziman (1979, p.119) pone en relieve que “hablar en rapidez como un atributo de una actividad técnica tan primitiva cuanto era la impresión por medio de la prensa manual […] cuya distribución se hacía en navíos con velas [...] puede que parezca un poco presumido”. Sin embargo, para los padrones de la época, ese cambio fue suficiente para que un número mucho más grande de científicos pudieran leer, discutir y escribir sobre las descubiertas

publicadas. Según Kronick (*apud* HARMON, 1992) el número de periódicos pasó de cuatro, en 1670, para ciento dieciocho en 1790.

Se hace importante acordar que el periódico fue, además, una respuesta al método científico que daba sus primeros pasos. Los resultados de las observaciones hechas aplicándose el método estaban bien representados en el formato de artículos cortos (HARMON, 1992).

El artículo científico se origina con el cambio de cartas entre científicos de diversos países europeos. Este método de comunicación, llamado de *Republique des Lettres,* se hizo tan eficiente que a las cartas se añadían comentarios de otros autores originado a otro texto, pudiendo ser bien diferente del original (SABBATINI, 1999).

Inicialmente, se escriban los artículos en forma de cartas enviadas para el editor, muchas de ellas, escritas en lengua vernácula, no en latín, con el propósito de alcanzar a un público más grande de científicos. Muchos periódicos mantienen, hasta hoy, un espacio para la publicación de cartas cuyo contenido pueda aludir a algún artículo publicado en el periódico o relatar el resultado de una investigación.

Sin embargo, con el paso de los años, la profesionalización y la especialización de la ciencia hizo con que los artículos y, consecuentemente, los periódicos sufrieran modificaciones (HARMON, 1992).

A partir del siglo XIX la estructura del artículo cambió. El rigor científico que se esperaba de un artículo no se combinaba con observaciones personales o lenguaje figurado: al escribir un artículo el autor debería limitarse a los hechos.

La estructura del artículo que se consolidó en este cambio fue la del artículo compuesto por:

a) Introducción, donde se delimita el problema de la investigación
b) Método, descripción de los materiales y métodos utilizados en la investigación.
c) Resultados, relato de los resultados obtenidos con la aplicación de los métodos de la investigación descriptos.
d) Discusión de los resultados.
e) Conclusión o Análisis del trabajo desarrollado.

Sólo después de 1850, los artículos empezaron a presentar lo que llamamos referencia bibliográfica, es decir, "[…] referencias explícitas a trabajos anteriores sobre los cuales se basa la nueva contribución […]" (PRICE, 1976, p.42)

Así, el primer libro de redacción científica lanzado en los EEUU es de 1927. Desde ese entonces, libros de temática semejante se vienen lanzando y la estructura del artículo arriba mencionado sigue siendo la que más se divulga y se acepta (HARMON, 1992).

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. Los periódicos científicos publicados en la Web pueden ser una herramienta cognitiva cuyas potencialidades aún no se evaluaron totalmente. Aunque publicados en la Web, periódicos electrónicos todavía son basados en el modelo tradicional en las publicaciones en papel y no se utiliza todo el potencial del medio electrónicos. Existen para que se los lean, se los evalúan y se los critique por personas; dependen 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 55% de los artículos publicados en periódicos indexados por el ISI, de 1981 hasta 1985, no recibieron ninguna citación por cinco años tras ser publicados. Y, aunque los artículos que fueron citados, no se lo hizo con frecuencia; solamente 42% de los artículos citados recibieron más que una citación.

De este modo, un artículo que demore a recibir citaciones puede formar parte de un grupo de artículos conocidos como de reconocimiento tardío (también llamado de descubierta prematura o descubierta 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).

Por su vez, Niiniluoto (2007, p.5), critica severamente al uso de los indicadores cienciométricos como instrumentos para la detección del progreso científico se dice "[…] que ellos hacen caso del *contenido semántico* de las publicaciones científicas".

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 con el conocimiento ya establecido (GARFIELD,1970).

El caso más exitoso de reconocimiento tardío es el artículo de Mendel sobre hibridación de plantas de plantas y publicado en 1865. El artículo fue citado pocas veces hasta ser "redescubierto" en 1900 (GARFIELD, 1970). Garfield fornece el ejemplo de otros cinco artículos que se puede considerar de reconocimiento tardío y que se identificó 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 relata más artículos que estarían en esta categoría, alzando algunas cuestiones pertinentes como:

> ¿El reconocimiento tardío es más prevalente en artículos metodológicos o conceptuales? […] ¿Existe alguna diferencia a lo largo de las últimas décadas, donde la existencia de mejores métodos para la recuperación de la información volvió aparentemente más difícil desconocer artículos relevantes? ¿O existe algún factor de retraso fundamental que debe inevitablemente afectar la aceptación de nuevas ideas a través del proceso educación-investigación? (GARFIELD, 1990, p.73).

En un último trabajo (GLÄNZEL; GARFIELD, 2004) los autores reafirman que los casos de artículos que tienen reconocimiento tardío son pocos y que la mayoría de los artículos importantes es mucho citada en los primeros tres a cinco años de publicación. De los 60 artículos reconocimiento tardío encontrados por ellos, 43% eran del área de ciencia de la vida.

Fue Van Raan (2004, p.467) quien llamó de "Bellas Durmientes" los artículos que "[…] no se perciben ('dormidos') a lo largo del tiempo y, entonces, casi que de repente, despiertan mucha atención ('son despertadas 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 en media por año fueron considerados "en sueño profundo", los que recibieron entre una y dos citaciones en media por año se consideró "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 en media. Por último, se consideró la “intensidad del despertar”; o sea, el número medio de citaciones cuatro años tras el “despertar”.

Entonces, de un universo de cerca de un millón de artículos, él encontró 41 artículos que después de un “sueño profundo” de diez años recibieron, en 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. Esto 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 con que los resultados y las conclusiones de investigación, como por ejemplo, los encontrados en los artículos científicos, puedan ser “interpretados” permitiendo que ordenadores puedan auxiliarnos en tareas más sofisticadas que demanden el procesamiento de dichos datos, disminuyendo la intervención humana y aumentándola precisión de las informaciones obtenidas. Particularmente 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 a otras bases de datos, limitando su utilidad.

Berners-Lee *et al*. (2001, p.35) propusieron el término Web Semántica y la definieron como:

> La Web semántica no se trata de una Web separada, pero una extensión de la actual, en la cual se utiliza la información con significado bien definido, aumentando la capacidad de los ordenadores para que trabajen en cooperación con las personas.

Para que la Web Semántica se vuelva una realidad se necesita que varias áreas del conocimiento trabajen cooperativamente. Una importante iniciativa en ese sentido es el *World Wide Web Consortium* (W3C). Creado en 1994, o W3C tiene el objetivo de “hacer con que la Web alcance todo su potencial a través del desarrollo de protocolos y directrices que garanticen el crecimiento a largo plazo de la Web”

Son fundamentales las ontologías para que la Web Semántica se vuelva en una realidad. Según Ding e Foo (2002, p.375): “Ontología es definida como una

especificación formal y explícita de una conceptuación compartida. Ella fornece un entendimiento compartido y común de un dominio que se puede comunicar a las personas y sistemas de aplicación".

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, p.19).

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

En esta 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 posee una estructura jerárquica, o *Metathesaurus* con cerca de 730.000 conceptos y más de 1 millón de nombres de conceptos. Él está complementado por una estructura clasificatoria llamada de *Semantic Network.*

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 *et al*., 1998).

El objetivo del UMLS es el de "[…] facilitar el desarrollo del sistemas de computadores que reaccionen como si "entendieran" el significado del lenguaje biomédico y de la salud" (NATIONAL…, 2008a, p.1). Para alcanzar este objetivo, la NLM produce y distribuye bases de datos de la UMLS, nombrados 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 informaciones que procesen, creen, recuperen, integren y/o agreguen datos y/o informaciones biomédicas y de la salud.

El UMLS se constituye de tres bases de conocimiento: un *Metathesaurus* que se forma por la agregación de más de cien vocabularios y clasificaciones; una

Semantic Network que "[…] provee una consistente clasificación de todos los conceptos representados en el *Metathesaurus*" (NATIONAL…, 2006, p.1) y un grupo de relaciones que existen entre conceptos del *Metathesaurus*; y el *Specialist* Lexicon que posee informaciones morfológicas , sintácticas y ortográficas para palabras del área biomédica y palabras frecuentemente usadas en inglés (NELSON; POWELL; HUMPHREYS, 2006). En este trabajo no trataremos del *Specialist* Lexicon.

El UMLS *Metathesaurus* es una amplia "[…] base de datos de vocabulario, de múltiples propósitos, multilingüe y que contiene informaciones sobre conceptos biomédicos y relacionados a la salud, sus varios nombres y sus relaciones entre ellos" (NATIONAL…, 2008b, p.1). Uno de esos vocabularios es el MeSH.

No obstante, uno de los aspectos que más generan polémica en la construcción del UMLS fue la definición de cómo el *Metathesaurus* debería ser elaborado. No se había acuerdo sobre la decisión de la NLM en construir el *Metathesaurus* a partir de la combinación de os 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 ULS (HUMPHREYS et al, 1998). La manera utilizada para la construcción del *Metathesaurus* implicó en que todos los conceptos, nombres y relaciones presentes en los diferentes vocabularios básicos estén presentes en el *Metathesaurus,* por ello

> […] cuando dos diferentes vocabularios fuente usan el mismo nombre para diferentes conceptos el Metathesaurus representa ambos significados y indica cual está presente en cual vocabulario fuente. Cuando el mismo concepto aparece en contextos jerárquicos diferentes, en diferentes vocabularios fuente, el Metathesaurus incluye todas las jerarquías. Cuando relaciones divergentes entre dos conceptos aparecen en diferentes vocabularios fuente, ambas las visiones incluidas en el Metethesaurus […] **el Metathesaurus no representa una abrangente ontología biomédica de autoría de la NLM o una única visión consistente del mundo (excepto en el más alto nivel de los tipos semánticos atribuidos a todos sus conceptos)** (NATIONAL…, 2008b, p.1, nuestra bastardilla).

La organización del *Metathesaurus* está hecha por conceptos y tiene por objetivo relacionar diferentes nombres para el mismo concepto de vocabularios diferentes. El *Metathesaurus* retiene todos los identificadores presentes en los vocabularios fuente, además de atribuir diversos tipos de identificadores permanentes y únicos para conceptos o significados, es llamado de *Concept Unique*

Identifier (CUI) y el no posee un significado propio aislado, es decir no se logra hacer ningún tipo de inferencia solamente leyéndolos.

En el caso de que se descubra que dos CUI se refieren a un mismo concepto, se remueve un CUI del Metathesaurus y toda la información relacionada al CUI retirado se transfiere para el CUI remanente. Un CUI que sea retirado nunca será reutilizado y la NLM mantiene un archivo que rastrea todos los cambios realizados en los CUI desde 1991.

Por fin, a Bondeireider (2001, p.4) "[…] el *Metathesaurus* puede proveer las bases para una ontología en el dominio biomédico".

3 PROCEDIMIENTOS METODOLÓGICOS

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, p.296) "[…] los científicos escriben en este formato, no solamente para cumplir los requerimientos de los periódicos, sino que también para atender las expectativas de la comunidad científica". Él también comenta que la mayoría de los manuales de redacción científica encoraja e 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, p.11) dice que la estructura IMRD "no es un formato arbitrario para la publicación, pero un reflejo directo del proceso de la descubierta científica".

En total, fueron analizados manualmente setenta y cinco artículos del área biomédica. Veinte artículos del periódico *Memórias del Instituto Oswaldo Cruz* (MIOC), veinte del *Brazilian Journal of Medical and Biological Research* (BJMBR), veinte artículos que trataban de la investigación con terapia génica germinal y quinze artículos de los ganadores del Lasker de 2006. El premio Lasker es un importante premio, otorgado anualmente y considerado tan importante cuanto el Nobel, pese que menos conocido. Él es considerado como una de las premiaciones que, por veces, anticipa 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. Ambos los

periódicos publican artículos en inglés y poseen un cuerpo editorial calificado, con revisores nacionales e internacionales.

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

En la búsqueda de artículos que trajeran indicios de descubiertas importantes, el tercer grupo de artículos analizados trataba de investigaciones sobre terapia genética geminal. La selección de os artículos de ese grupo se hizo a través de la lectura de tres artículos de recién repaso del área (NATIONAL…, 2006; FRIEL; SAR; MEE, 2005; BONGSO; RICHARDS, 2004) en el que presentaban una visión histórica de la investigación de la terapia génica germinal, resaltando los avances más importantes, informaciones extremamente relevantes para esa investigación.

Aún buscando artículos que se reportaran a descubiertas importantes, un último grupo de artículos fue elegido. Ese grupo fue compuesto de artículos que constaban de 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 a la descubierta de la telomerase. Cada autor laureado proveyó un listín de sus trabajos que creían más importantes y, de la unión de los tres listines, se obtuvo los 15 artículos analizados. De este último constan artículos de varios periódicos científicos como *Cell* y *Nature,* periódicos con alto factor de impacto, 29,887 y 28,751, respectivamente.

El análisis del contenido de los artículos del Laskerfue facilitada por 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; GREIDER; SZOSTAK, 2006) por ocasión de la premiación con Lasker. En ella, los autores presentan la trayectoria de la investigación resaltando los artículos que creían más importantes y especificando la contribución dada por cada uno de ellos.

Debido a la estructura textual altamente formalizada de sus artículos, se seleccionaran periódicos del área Biomédico. La mayor parte de los artículos del MIOC era del área de la microbiología; los del BJMBR eran más heterogéneos habiendo una predominancia de artículos de las áreas de fisiología y neurociencias; por fin, los artículos de terapia génica germinal y telomerase que trataban de cuestiones relacionadas a la genética. Es importante enfatizar que la elección de esos periódicos no se hizo en un único momento, pero, gradualmente, a lo largo del desarrollo de la investigación, buscando siempre por artículos que trajeran descubiertas científicas 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 un consecuente (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; SZOSTAK, 1989) la conclusión fue sintetizada en la siguiente afirmación: El acortamiento del telómero causa senescencia celular.

O esquemáticamente:

Antecedente: acortamiento del telómero

Relación: causa

Consecuente: senescencia celular

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

El artículo al ser publicado es casi que inmediatamente indexado. Considerando que la indexación fue hecha con los mejores términos disponibles en la época de publicación y establecidos el antecedente, la relación y el consecuente, se verificó en qué grado estos elementos estaban representados en la indexación

MeSH del artículo. A eso, damos el nombre del mapeado. Si todos los elementos (el antecedente, relación y consecuente) fueran mapeados en el UMLS, el artículo era considerado completamente mapeado. Se uno o dos elementos no fueran mapeado en el UMLS, el artículo era considerado no mapeado.

Aunque “novedad científica” sea una noción sin una definición exacta y la literatura muestre intentos de definición más calificativa, 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 públicas en el mismo dominio científico del artículo.

4 RESULTADOS FINALES

De entre los 75 artículos analizados, el grupo que reportaba indicios de descubiertas importantes, es decir, los artículos de los ganadores del premio Lasker de 2006, seguidos de los artículos de terapia génica germinal – obtuvo el peor porcentaje de mapeado. En ese grupo, o no se mapeó todos los artículos o se mapeó parcialmente. De entre los artículos mapeados parcialmente, el mapeado sólo se hizo posible a causa del tipo de relación. No se permitió mapear ni el antecedente ni el consecuente.

En el grupo de artículos sobre terapia génica germinal un 80% se mapeó parcialmente y un 20% no se mapeó.

Al sumar los artículos del MIOC y BJMBR y comparándolos con la suma de los artículos del Lasker y de terapia génica germinal (L+CT), los artículos MIOC+BJMBR recibieron mapeado completo en porcentaje mayor (25%) que los L+CT (0%).

Tabla I – Análisis de la representación de la conclusión de los artículos de los periódicos Memórias do Instituto Oswaldo Cruz (MIOC), Brazilian Journal of Medical and Biological Research (BJMBR), artículos relevantes sobre terapia génica germinal y artículos sobre la investigación de la telomerase de ganadores del premio Lasker.

Artículos Analizados	MIOC	BJMBR	Terapia Génica Germinal	Telomerase	TOTAL
Totalmente mapeados	7 (35%)	3 (15%)	0 (0%)	0 (0%)	10
Parcialmente mapeados	13 (65%)	11 (55%)	16 (80%)	6 (40%)	44
No mapeados	0 (0%)	6 (30%)	4 (20%)	9 (60%)	21
Total de artículos	20	20	20	15	75

5 CONSIDERACIONES FINALES

Se observa, en la literatura, que el término “ontología biomédica” puede referirse tanto a terminologías usadas para indexar la literatura científica como a las ontologías computacionales formales de alto nivel. Su desarrollo, evolución e integración es un trabajo científico complejo. Mientras, por ejemplo, la *Gene Ontology* fue desarrollada recientemente para permitir que se comparta una terminología común para los apuntes de productos genéticos, otras, con el MeSH, que forma parte del UMLS, tiene que lidiar con el legado de millones de registros indexados en base de datos bibliográficos como el PubMed y el Medline (BODENREIDER, 2008).

Se puede suponer que, al volverse más formales, las terminologías biomédicas están evolucionando para ser bases del conocimiento. Las ontologías están clasificadas según su grado de formalismo, se puede variar desde una simple taxonomía usada por personas hasta una ontología altamente formalizada codificada en un lenguaje como OWL.

El objetivo de este trabajo fue demostrar la viabilidad de un método que permite comparar las conclusiones de artículos científicos con el conocimiento registrado en ontologías públicas en la Web a fin de identificar posibles descubiertas 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 descubiertas científicas de manera más precisa. Se usó aquí el MeSH como una herramienta en la falta, de momento, de otra mejor.

Los resultados indican que el grado de éxito/ no éxito por el mapeado de la representación de la conclusión MeSH se puede asociar al hecho de que los artículos relaten descubiertas científicas importantes. Parece metodológicamente posible proponer un procedimiento en el que los autores expresen su principal conclusión de manera sintética y que, la misma, sea automáticamente procesada y comparada al conocimiento científico ya previamente establecido y representado en las ontologías públicas.

La creciente cantidad de artículos que se publica 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. Un método automático que pueda apuntar indicios de novedad puede optimizar este proceso, para que la atención del investigador o del gestor de C&T pueda encontrarse en 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 se están dando par la ciencia. La indexación se hace, posteriormente y pronto la publicación, cuando los artículos son incluidos en bases de datos o repositorios como el Medline o PubMed.

De esta manera, una nueva descubierta científica puede crear nuevos conceptos a los cuales un término que aun no haya sido cuñado en las bases de datos terminológicos como el UMLS. Así, existe un retraso entre la descubierta de un fenómeno, o concepto y la actualización del UMLS. Como ya citado anteriormente, el término telomerase fue relatado en 1985 (GREIDER; BLACKBURN, 1985), pero sólo se lo incorporó al MeSH en 1995, diez años después.

Hace falta poner en relieve que, en algunos casos la “novedad científica” no está acompañada de la creación de nuevos términos, pero, 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 se están evolucionando para una mayor formalización y necesitarán de nuevos métodos de curaduría (WILLIAMS; 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, tras totalmente automatizado y implementado, pueda volverse en más una herramienta de evaluación de la producción científica, complementar a los ya tradicionales métodos bibliográficos y cienciométricos.

REFERÊNCIAS

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific American**, v.284, n.5, p.34-43, 2001. Disponible en: <http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>. Acceso en: 16 jun. 2006.

BLACKBURN, E. H.; GREIDER, C. W.; SZOSTAK, J. W. Telomeres and telomerase: the path from maize, *Tetrahymena* and yeast to human cancer and aging. **Nature Medicine**, v.12, n.10, p.vii-xii, 2006.

BONDEREIDER, O. Medical ontology research. **Report to the board of scientific counselors of the Lister Hill National Center for Biomedical Communications**, 2001. Disponible en: <http://mor.nlm.nih.gov/>. Acceso en: 06. ene. 2008.

_____. Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. **Yearb. Med. Inform.**, p.67-79, 2008. Disponible en: <http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=2592252&blobtype=pdf>. Acceso en: 31 ago. 2009.

BONGSO, A.; RICHARDS, M. History and perspective of stem cell research. **Best Practice & Research Clinical Obstetrics & Gynaecology**, v.18, n.6, p.827-842, 2004.

BURROUGH-BOENISH, J. International reading strategies for IMRD articles. **Written Communication**, v.16, n.3, p.296-316, 1999.

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

DE ROURE, D.; JENNINGS, N.; SHADBOLT, N. Research agenda for the Semantic Grid: a future s-Science infrastructure. **Report Commissioned for EPSRC/DTI Core e-Science Programme**, 2001. 78p.

DING, Y.; FOO, S. Ontology research and development - Part 2: A review of ontology mapping and evolving. **Journal of Information Science**, v.28, n.5, p.375-88, 2002.

FRIEL, R.; SAR, S.; MEE, P. Embryonic stem cells: understanding their history, cell biology and signalling. **Advanced Drug Delivery Reviews**, v.57, n.13, p.1894-1903, 2005.

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

_____. More delayed recognition - Part 2: From inhibit to Scanning Electron Microscopy. **Essays of an Information Scientist**, v.13, p.68-74, 1990.

GLÄNZEL, W.; GARFIELD, E. The myth of delayed recognition. **The Scientist**, v.18, n.11, p.8, 2004.

GREIDER, C. W.; BLACKBURN, E. H. Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. **Cell**, v.43, p.405-413, 1985.

HAMILTON, D. P. Publishing by – and for? – the numbers. **Science**, v.250, n.4986, p.1331-32, 1990.

HARMON, J. E. Evolution of the Scientific Paper. In: INTERNATIONAL PROFESIONAL COMMUNICATION CONFERENCE (IPCC), 1992, Santa Fé. **Proceedings…** [S.I.: s.n.], 1992. p.468-475

HUMPHREYS, B. L. *et al*. The Unified Medical Language System: an informatics research collaboration. **Journal of the American Medical Informatics Association**, v.5, n.1, p.1-11, 1998.

INTERNATIONAL COMMITTEE OF MEDICAL JOURNAL EDITOR (ICMJE). Uniform requirements for manuscripts submitted to Biomedical journals: writing and editing for biomedical publication. ICMJE, 2008. p.1-16. Disponible en: <http://www.icmje.org/#prepare>. Acceso en: 22 oct. 2008.

JACOB, E. K. Ontologies and the Semantic Web. **Bulletin of the American Society for Information Science and Technology**, v.29, n.4, p.19-22, 2003.

HARMON, J. E. Evolution of the Scientific Paper. In: INTERNATIONAL PROFESIONAL COMMUNICATION CONFERENCE (IPCC), 1992, Santa Fé. **Proceedings…** [S.I.:s.n.], 1992. p. 468-475

KUHN, T. S. **A estrutura das revoluções científicas**. São Paulo: Perspectiva, 2005. (Coleção Debates, 115)

LUNDBLAD, V.; SZOSTAK, J. W. A mutant with a defect in telomere elongation leads to senescence in yeast. **Cell**, v.57, n.4, p.633-643, 1989.

MARCONDES, C. H. *et al*. Ontological and conceptual bases for a scientific knowledge model in biomedical articles. **RECIIS**, v.3, n.1, p.19-30, 2009. Disponible en: <http://www.reciis.cict.fiocruz.br/index.php/reciis/article/view/240/251>. Acceso en: 8 abr. 2009.

MEADOWS, A. J. **A Comunicação científica**. Brasília, DF: Briquet de Lemos, 1999. 268p.

NATIONAL INSTITUTES OF HEALTH. **The human embryonic stem cell and the human embryonic germ cell**. Disponible en: <http://stemcells.nih.gov/>. Acceso en: 8 mar. 2006.

NATIONAL LIBRARY OF MEDICINE. **Unified Medical Language System**: Fact sheet. 2006. Disponible en: <http://www.nlm.nih.gov/pubs/factsheets/umls.html>. Acceso en: 04 ene. 2008.

_____. **Unified Medical Language System**. 2008a. Disponible en: <http://www.nlm.nih.gov/research/umls/about_umls.html>. Acceso en: 04 ene. 2008.

_____. **Unified Medical Language System - Metathesaurus**, 2008b. Disponible en: <http://www.nlm.nih.gov/research/umls/meta2.html>. Acceso en: 04 ene. 2008

NELSON, S. J.; POWELL, T.; HUMPHREYS, B. L. **The Unified Medical System® (UMLS®) project**, 2006. Disponible en: <http://www.nlm.nih.gov/mesh/umlsforelis.html>. Acceso en: 04 ene. 2008.

NIINILUOTO, I. Scientific Progress. In: ZALTA, E.N. (Ed.). **The Stanford Encyclopedia of Philosophy**, Feb. 2007. Disponible en: <http://plato.stanford.edu/archives/fall2008/entries/scientific-progress/>. Acceso en: 01 feb. 2008.

OLDENBURG, H. Introduction. **Philosophical Transactions**, v.1, n.1, p.1-2, 1665.

PRICE, D. J. de S. **O desenvolvimento da ciência**. Rio de Janeiro: Livros Técnicos e Científicos, 1976. 96p.

SABBATINI, M. **Evolución histórica de las publicaciones científicas: de la Republique des Lettres hasta la World Wide Web**. Salamanca, 1999. Trabalho de curso apresentado ao Máster CTS, Cultura y Comuncación en Ciencia y Tecnología. Disponible en: <http://www.sabbatini.com/marcelo/artigos/1999sabbatini-republique.pdf>. Acceso en: 02 jun. 2008.

SMITH, B. Ontology (Science). **Nature Precedings**, 2008. Disponible en: <http://hdl.handle.net/10101/npre.2008.2027.2>. Acceso en: 1 ago. 2009.

STEIN, L. D. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. **Nature Reviews Genetic**, v.9, p.678-688, Sept. 2008.

VAN RAAN, A. F. J. Sleeping Beauties in Science. **Scientometrics**, v.59, n.3, p.467-472, 2004.

WILLIAMS, J.; ANDERSON, W. Bringing ontology to the Gene Ontology. **Comparative and Functional Genomics**, v.4, p.90-93, 2003. Disponible en: <http://hindawi.com/GetPDF.aspx?doi=10.1002/cfg.253>. Acceso en: 31 jul. 2009.

ZIMAN, J. **Conhecimento público**. Belo Horizonte: Itatiaia; São Paulo: EDUSP, 1979. (Coleção o Homem e a Ciência).