REVISTA ESPAÑOLA DE DOCUMENTACIÓN CIENTÍFICA
29, 2, ABRIL-JUNIO, 220-238, 2006
ISSN 0210-0614

CONSTRUCCIÓN DE UNA METODOLOGÍA PARA IDENTIFICAR INVESTIGADORES MEXICANOS EN BASES DE DATOS DE ISI

C. A. Macías-Chapula*, J. A. Mendoza-Guerrero, I. P. Rodea-Castro, A. Gutiérrez-Carrasco

Resumen: La evaluación de la ciencia es importante para apoyar los procesos de decisión y de gestión en materia de política científica. Visto como un sistema, los resultados de la actividad científica (indicadores de *output*), se pueden evaluar a través de estudios bibliométricos, cienciométricos y webométricos. En general estos estudios se realizan utilizando las bases de datos del *Institute for Scientific Information* (ISI). Uno de los problemas mayores con el uso de estas bases de datos es el manejo de nombres de autores, principalmente hispanos. El propósito de este trabajo fue el identificar la cobertura de los nombres de los miembros del Sistema Nacional de Investigadores (SNI) de México en la base de datos *National Citation Report*-México (NCR), de ISI. El objetivo final fue el de construir una metodología que ayude a incrementar la recuperación y precisión en la validez de los datos confrontados del SNI, en NCR. El estudio se realizó en dos fases. La primera consistió en identificar la cobertura de los 9,201 miembros del SNI en NCR, para el periodo 1984-2002. Para la segunda fase se seleccionó una muestra de 658 nombres, y se realizó una búsqueda exhaustiva de autores en NCR, incluyendo criterios de validación. Los resultados ayudaron a construir una metodología que culminó en el agrupamiento de cuatro categorías de nombres con diferentes niveles de dificultad y certeza en la recuperación de registros en NCR. Con esta metodología se logró incrementar hasta en 26,9% la recuperación de registros. El documento presenta el modelo conceptual de esta metodología emergente; describe las líneas de investigación a seguir y discute sobre las implicaciones de este tipo de estudios.

Palabras clave: base de datos, bibliografía, estudios bibliométricos, información, almacenamiento y recuperación, lenguaje, método, Sistema Nacional de Investigadores/México, NCR-México.

Abstract: Indicators of science performance and evaluation are important to support decision making processes in science policy. In this context, *output* science indicators are analyzed through bibliometric, scientometric and webometric studies, usually conducted in databases produced by the Institute for Scientific Information. (ISI). One of the major problems with

* Hospital General de México. México, D.F. Correo-e: cesarmch@liceaga.facmed.unam.mx.
Recibido: 15-8-2005; 2.ª versión: 4-5-2006.

the use of these products however, is related to the variations of a given name in the author field. This is particularly relevant in the case of hispanic names, where a search strategy needs to consider significant variations to an author name. Proposals however have been limited to being aware of the situation and to submit recommendations to database producers, journal editors and even authors. Up to date no reports have been published on the method or approach used to analyze and solve this problem. The purpose of this work was to identify the coverage of the members of Mexico's Researchers National System (Sistema Nacional de Investigadores, SNI) in ISI's National Citation Report-Mexico (NCR) data base. The final goal was to construct a methodology so as to increase recall and precision rates in the coverage of SNI members in NCR. The study considered two phases. Phase one lead to the identification of the 9,201 SNI members in NCR for the period 1984-2002. In the second phase, a sample of 658 names was selected from SNI members. And an exhaustive search of author names was conducted, including precision criteria such as validation. Results helped to construct a methodology that lead to the grouping of four categories of names, each with a different level of difficulty and precision in the recall of data from NCR. An increase of up to 26.9% in the recall ratio from NCR was obtained through the use of this methodology. This document describes the conceptual model that emerged from the methodology, and discusses the research lines to follow as well as the implications of the study.

Keywords: databases, bibliography, bibliometrics, information storage and retrieval, language, Mexico/Researchers National System, National Citation Reports-Mexico, method.

1. Introducción

La evaluación del desempeño de la actividad científica de un individuo, un grupo de investigadores, una institución o un país, es tarea obligada en todo proceso de gestión relacionado con la organización y administración de los recursos asignados a la investigación y el desarrollo. Los resultados obtenidos con la evaluación apoyan a la vez los procesos de toma de decisión en materia de política científica. Cuando la evaluación se realiza a nivel nacional y se analiza como sistema, se vuelve necesario identificar no sólo los indicadores de *input*, como gasto en ciencia, número de investigadores, infraestructura, etc., sino también los indicadores de *output*, como productividad, impacto, y/o *benchmarking*, tal como lo han venido realizando organismos internacionales establecidos para ese propósito (National Science Board, 2004; European Commission, 2003; OCD, 2002; RICYT, 2002).

1.2. Las bases de datos del Institute for Scientific Information (ISI)

El interés por evaluar el desempeño de la actividad científica a través de los indicadores de *output*, ha permitido la abundante conducción de estudios bibliométricos, cienciométricos y, recientemente, webométricos, en un intento por obtener indicadores cuantitativos que complementen la evaluación cualitativa de la producción científica (Small, 2003; Aguillo, 2001; Macías-Chapula, 2002; 1994; 1991; Bordons y Zulueta, 1999; Small, 1999; Macías-Chapula y cols., 1998; Van-Raan, 1993; Sancho, 1990; Velho, 1990; Moed, 1989; Moravsik, 1989). En la conducción de este tipo de estudios, se vuelve imprescindible el uso de las bases de datos del *Institute for Scientific Information* (ISI). Desde su origen, los productos del ISI se han utilizado como herramientas importantes para evaluar el desempeño de la actividad científica de países, instituciones, grupos de investigación y hasta individuos (Garfield y Welljams, 1992; Garfield, 1990; 1979). De todas las bases de datos del ISI, el *Science Citation Index* (SCI), por su carácter multidisciplinario, es quizá la más utilizada para conducir estudios métricos de la ciencia. El NCR-México es una base de datos electrónica solicitada por México al ISI; y contiene los datos de artículos publicados por autores e instituciones mexicanas durante el periodo 1981-2002. Además de contener la información bibliográfica de cada uno de los artículos indizados por ISI, la base de datos integra el número total de citas recibidas por cada uno de los artículos (ISI, 2003).

El uso de las bases de datos del ISI para evaluar la actividad científica ha sido igualmente objeto de crítica; ello debido principalmente a la ausencia de una teoría establecida sobre el análisis de citas (Cronin, 1984) y a la baja representación en esas bases de datos, de revistas científicas generadas por los países de la llamada *periferia* (Zulueta y Bordons, 1999; Gómez y Bordons, 1996). Salvo contados proyectos como SciELO (BIREME, 2004) y Latindex (Cetto y Alonso, 1998), no se han realizado esfuerzos por mejorar esta situación. Ello ha tenido repercusiones serias en países en vía de desarrollo, donde los científicos se ven obligados a publicar en revistas que estén incluidas en las bases de datos del ISI, la mayoría en idioma inglés. Esta inclusión de los investigadores en ISI, les otorga la visibilidad internacional buscada a través de las citas para sus trabajos y, por lo tanto, su impacto.

En la literatura se han encontrado estudios sobre las limitaciones en el uso de las bases de ISI (Bordons y Zulueta, 1999; MacRoberts y MacRoberts, 1996; Hamilton, 1991). El manejo, por ejemplo, de nombres de investigadores se debe hacer con sumo cuidado. Al no contar con un criterio homogéneo de búsqueda de nombre de autores, existe el riesgo de confundir nombres, omitir siglas; o bien, asignar producción, impacto y desempeño equivocados a investigadores o grupos específicos. Esto es particularmente relevante en el manejo de nombres hispanos (Esteve-Fernández, 2003; Ruiz-Pérez y cols., 2002; Gómez y cols., 1997; López-Cózar, 1997). Sin embargo, existen trabajos que presentan aproximaciones metodológicas

para el tratamiento del problema de normalización del campo de autor en estas bases de datos (Torvik y cols., 2005; Costas y Bordons, 2005; Wooding y cols., 2004; Costas-Comezaña y García-Zorita, 2003). A pesar de que esta situación de alerta es conocida por la mayoría de los investigadores dedicados a esta área del conocimiento, no se ha presentado en la literatura una metodología o enfoque que ayude a establecer procedimientos para organizar y validar registros de las bases de datos de ISI, adecuadamente. En efecto, la mayoría de los trabajos discuten los problemas, implicaciones y las limitaciones que fluctúan alrededor de los nombres de los autores de publicaciones; refieren propuestas de alerta a esta situación y emiten recomendaciones a diferentes instancias (Castro y cols., 2004; Scoville y cols., 2003; Ruiz-Pérez y cols., 2002; Siebers y Holt, 2000; Reyes y cols., 2000; Flanagin y cols., 1998; Wilcox, 1998; Drenth, 1998; Rennie y cols., 1997; Epstein, 1993). No aportan, sin embargo, mayor análisis para construir una metodología que aborde soluciones al problema relacionado con el manejo de los nombres de autores. La construcción de una metodología se vuelve particularmente importante en aquellos casos donde se manejan grandes cantidades de registros; o bien, no se cuenta con el acceso a los currícula o contacto con los investigadores.

1.2. El Sistema Nacional de Investigadores de México (SNI)

En la región latinoamericana se han orquestado al interior de algunos países, el uso de instrumentos de evaluación y estímulos al desempeño de la actividad científica de sus investigadores en un intento por mejorar la productividad y el rendimiento nacionales. En México, el Consejo Nacional de Ciencia y Tecnología (CONACyT) creó, desde julio de 1984, el Sistema Nacional de Investigadores (SNI), como una iniciativa que permitiría distribuir recursos financieros adicionales a los científicos, sin recurrir a los incrementos generalizados de salario (*Diario Oficial de la Federación*, 1984). Ésta fue la primera medida del gobierno federal que, a nivel nacional, condujo una evaluación del desempeño individual de los investigadores del país. Esta iniciativa permitió, además de identificar quién estaba en condiciones de recibir el nombramiento de investigador a nivel nacional, asignar un estimulo económico mensual basado en el desempeño y el nivel de cada uno de los investigadores dentro del sistema. A la fecha se reconocen cuatro niveles del SNI; de menor a mayor jerarquía éstos fluctúan desde el *candidato* a investigador nacional, hasta el investigador nacional, *nivel III*. El SNI inició con un total de 1,396 investigadores en 1984. Para el 2003, contaba con un total de 9,201 miembros; un crecimiento de aproximadamente siete veces desde su creación.

Todo miembro del SNI, dependiendo del nivel que ocupe, se asume como un investigador en activo, de tiempo completo, formador de recursos humanos, líder de proyectos de investigación y con suficiente productividad e impacto internacional.

Se espera que dicha productividad tenga una visibilidad en fuentes secundarias convencionales de información; y que el impacto sea establecido en base a las citas recibidas a sus trabajos publicados en la corriente internacional; principalmente reflejadas en los productos del ISI.

En la literatura se han publicado diversos estudios sobre la representación de México en las bases de datos del ISI (Almeida, 2003; Collazo-Reyes y Luna-Morales 2002; Licea y Santillán-Rivero 2002; Russell, 1998; Macías-Chapula y Rodea-Castro, 1997; Pellegrini y Goldbaum, 1997; Macías-Chapula 1995). Dichos estudios, sin embargo, han tomado el universo de la producción del país; o bien, se han conducido para identificar la representación del país en diversas áreas del conocimiento, instituciones o grupos de investigación. Ningún estudio se ha realizado sobre la representación de los miembros del SNI de México, en las bases de datos del ISI. Se asume que, dados los criterios de evaluación del SNI, la producción de la gran mayoría si no es que la totalidad de los miembros del sistema, se encuentren incluidos en los productos del ISI. Esta inquietud por obtener dicha representación fue la que llevó a la conducción original del estudio.

2. Propósito

El objetivo de este trabajo es el de presentar los resultados preliminares de un proyecto de investigación sobre la cobertura de los miembros del Sistema Nacional de Investigadores (SNI) de México, en la base de datos *National Citation Report-México* (NCR), del ISI. El estudio pretende construir una metodología que ayude a incrementar la recuperación y precisión de los nombres de los investigadores incluidos en las bases de datos del ISI.

3. Método

Para la conducción del estudio se utilizaron dos bases de datos en soporte magnético; una del SNI, México, correspondiente a los nombres de los investigadores miembros del SNI vigentes al periodo 1984-2003; y otra correspondiente a los nombres de los investigadores que aparecen en NCR de ISI, con una cobertura de 1981 a 2002. De los registros correspondientes a la base de datos del SNI, se clasificaron los datos conforme a los niveles del sistema para obtener la distribución de la cantidad de los investigadores por nivel. Esto es, candidatos a investigador nacional; investigador nacional nivel I; nivel II y nivel III. Se diseñaron dos fases de análisis para obtener la correspondencia de los registros entre ambas bases de datos. Ello con la intención de obtener resultados comparativos en dos niveles de resolución diferente.

Las fases del estudio fueron las siguientes:

– *Fase 1,* correspondencia de nombres del SNI en sus cuatro niveles, la base de datos NCR, sin profundizar en el análisis exhaustivo o la validación de los datos. El propósito de esta fase fue el de identificar en un primer nivel de resolución, el universo de investigadores, su cobertura y distribución entre ambas bases de datos, para un mismo período de análisis (1984-2002).
– *Fase 2,* correspondencia de nombres del SNI-Nivel III, con la base de datos NCR, conduciendo un análisis profundo y exhaustivo de los datos. El propósito de esta fase fue el de explorar en un segundo nivel de resolución, el enfoque metodológico que ayudara a incrementar la recuperación y la precisión en la correspondencia de los registros de ambas bases de datos.

Para la fase 1, se procedió a identificar quiénes y cuántos investigadores, pertenecientes al SNI, se encontraban incluidos en la base de datos NCR, esto se realizó a través de la confrontación de la listas de los nombres de los investigadores entre ambas bases (NCR y de SNI), los nombres fácilmente identificados fueron aquellos que se encontraron escritos de la misma forma en ambas bases, es decir aquellos que coincidieron con los apellidos completos y las siglas o sigla del primer y segundo nombre de pila, estos casos fueron la minoría. Para algunas otras variantes se identificaron algunos de los nombres de investigadores apoyándose en el área de conocimiento.

Inicialmente se trabajó con un listado de 9,201 nombres registrados en el SNI, el año de 2002. Sin pretender ser exhaustivo, este primer análisis ayudó a identificar las diversas formas en que un nombre es registrado en el NCR, las homonimias y el grado de dificultad para validar los registros recuperados. El SNI por ejemplo, despliega los nombres completos de los investigadores, con apellidos paterno, materno y nombre(s) de pila. El NCR por otro lado, en la mayoría de los casos registra únicamente el apellido paterno y las siglas de el o los nombres de pila, o bien con las siglas del segundo apellido y del nombre de pila. De esta forma se identificaron variaciones importantes en los nombres y su dificultad para identificarlos. Otra variante encontrada en NCR fue la unión del apellido paterno y materno, con las siglas de los nombres de pila. La experiencia lograda en el análisis de los nombres en esta primera fase, ayudó a estructurar el enfoque a seguir en la segunda fase.

Para la segunda fase, el universo de análisis se limitó a los 658 nombres de los investigadores miembros del SNI-*Nivel III*. Las razones por las cuales se seleccionó esta muestra fueron las siguientes: (a) los investigadores de este nivel son los que tienen valores más altos de producción e impacto; (b) los investigadores de este nivel tienen mayor probabilidad de encontrarse representados en NCR y (c) el nivel III del SNI es el que cuenta con menor cantidad de miembros; equivalente al 7,15% del total.

En esta segunda fase, el período de estudio fue de 1984-2002 y se limitó a la búsqueda de investigadores en NCR, correspondientes al nivel III del SNI. De primera instancia, se seleccionó el nombre del investigador con sus dos apellidos y siglas de, el o los nombres de pila. En caso de no encontrar registro alguno, se procedió a seleccionar el apellido paterno y las siglas de, el o los nombres de pila. De esta forma, se manejaron todas las permutaciones posibles del nombre, aclarando que en el caso de investigadores mexicanos regularmente asientan bajo el apellido paterno y en muy escasas ocasiones bajo el apellido materno. Con este procedimiento, NCR proporcionó todos los apellidos y siglas que correspondían al nombre de referencia del SNI. Una vez seleccionados los apellidos y siglas de los nombres de pila, se recuperaron para cada nombre –con sus variantes– la producción de artículos y citas a los trabajos correspondientes.

Para validar la confiabilidad de los datos, por ejemplo, en caso de homonimia, se procedió a identificar la correspondencia de los nombres con el área del conocimiento que asignó NCR a sus registros. Sin embargo de existir una homonimia y una misma área de conocimiento sería necesario incluir otros procedimientos de validación. Por ejemplo se consideró, aunque *no* incluidos en este estudio, la consulta al currículum vitae de los investigadores y la comunicación personal con ellos, para confirmar los resultados obtenidos. Esta metodología de recuperación de producción científica, en un futuro se pretende incorporar a la estrategia de recuperación de la producción generada por el Hospital General de México, con la finalidad de obtener su visibilidad y posicionamiento en el sector salud de México a través de los años.

4. Resultados

Para la *fase 1* del estudio se encontró que sólo 57.96% del total de los miembros del SNI de México aparecían incluidos en la base de datos del NCR. La tabla I presenta la distribución de los miembros del SNI por nivel, y su cobertura correspondiente en NCR. En esta primera fase del estudio se detectaron diversas discrepancias en la forma en que se despliegan los datos de los nombres para ambas bases de datos. En la base de datos del SNI aparecen los apellidos y nombre(s) de pila completos. Sin embargo, en la base del NCR aparecen hasta siete variantes principales en los nombres. El Anexo 1 sintetiza las variantes encontradas para ambas bases de datos.

Además de las variantes de los nombres de los investigadores, el NCR redunda en homonimias al seleccionar apellidos y siglas similares en caracteres, pero correspondientes a investigadores diferentes. La tabla II describe, por ejemplo, el caso de varios investigadores reconocidos por NCR con los mismos caracteres, pero que en realidad corresponden a investigadores diferentes. Esto se validó con la identificación del área del conocimiento a la que pertenece cada uno de ellos en NCR.

Tabla I
Distribución de los investigadores pertenecientes al Sistema Nacional de Investigadores (SNI) de México, en función de su nivel y cobertura en el *National Citation Report* (NCR) del *Institute for Scientific Information* (ISI).

*Nivel en el SNI**	*Número*	*%*	*Núm. de registros localizados en NCR***	*%*
Candidatos	1.582	17,19	780	49,30
Nivel I	5.374	58,41	3.100	57,69
Nivel II	1.587	17,25	1.018	64,15
Nivel III	658	07,15	435	66,11
Total	**9.201**	**100,00**	**5.333**	**57,96**

* Cifras correspondientes al año 2003.
** Cifras correspondientes al periodo 1981-2002.

Tabla II
Identificación de investigadores del SNI que reflejan homonimia en NCR y que finalmente son validados a través del área del conocimiento a la que pertenecen

Despliegue del nombre del investigador en el SNI	*Despliegue del nombre del investigador en NCR*	*Área del conocimiento a la que pertenece el investigador en NCR*
ALBA ANDRADE FERNANDO	ALBA F	FÍSICA
ALBA HERNÁNDEZ FRANCISCO	ALBA F	MEDICINA
BÁEZ PEDRAJO ARMANDO	BAEZ A BAEZ AP	QUÍMICA ATMOSFÉRICA QUÍMICA ATMOSFÉRICA
LÓPEZ CASTRO GABRIEL	LOPEZ G	FÍSICA
LÓPEZ CARRANZA GREGORIO	LOPEZ GC	PSICOLOGÍA

El proceso de análisis de los nombres, con sus permutaciones y diferentes niveles de precisión obtenidos en la primera fase del estudio, indicó una falta de homogeneización de los nombres en NCR. Este hallazgo ayudó a establecer para la *segunda fase* del estudio, un enfoque organizacional en el manejo de los registros, de tal manera que, ante la falta de homogeneidad en los nombres, se pudiera incrementar

la recuperación y la certeza de los registros de NCR. De esta forma, como resultado del estudio, se obtuvo una jerarquía de cuatro categorías de grupos de nombres. Estas categorías ayudaron a clasificar los registros en base a dos criterios: uno de precisión y otro de validación. En este contexto, las categorías fluctuaron de una mayor a una menor precisión en la recuperación de registros; y de una menor a una mayor dificultad en la recuperación y validación de los resultados.

Las categorías que emergieron como resultado del análisis fueron las siguientes:

- *Categoría 1.* Nombres con mínima dificultad de recuperación; mayor precisión. Aquí se incluyeron aquellos apellidos y siglas de nombres de pila que coincidieron sin mayor esfuerzo de validación. En el caso de homonimias, los datos se validaron a través de la identificación de las áreas del conocimiento a las que pertenecieron los registros. La tabla III presenta un ejemplo de los nombres que integraron esta categoría.

Tabla III
Categoría 1. Nombres con mínima dificultad de recuperación y mayor precisión en la validación de los datos que se corresponden entre SNI y NCR

Apellido y nombre en el SNI-III	*Datos en NCR*	*Citas*	*Artículos*
ACEVES RUIZ JORGE	ACEVES J	515	42
ALUJA SCHUNEMAN MARTÍN	ALUJA M	235	42
ARÉCHIGA URTUZUASTEGUI HUGO	ARECHIGA H	195	32
BERMÚDEZ RATTONI FEDERICO	BERMUDEZ RATTONI F	395	43
BEYER FLORES CARLOS JOSÉ	BEYER C	364	36

- *Categoría 2.* Nombres con mediana dificultad de recuperación; regular precisión. Aquí se incluyeron aquellos nombres que requirieron la búsqueda de dos a cuatro formas diferentes, a través de las permutaciones de sus apellidos y siglas de nombres de pila. En algunos casos hubo que validar homónimos con las áreas del conocimiento correspondiente. En caso necesario, la consulta del currículum vitae de los investigadores ayudaría también a la validación de los datos para incrementar la precisión de los resultados. La tabla IV presenta un ejemplo de los nombres que integraron esta categoría.

Tabla IV
Categoría 2. Nombres con mediana dificultad de recuperación y regular precisión en la validación de los datos que se corresponden entre SNI y NCR

Apellido y nombre en el SNI-III	*Datos en NCR*	*Citas*	*Artículos*
GÓMEZ PUYOU ARMANDO	GOMEZ PUYOU A	329	48
	GOMEZ A	317	83
SEPÚLVEDA AMOR JAIME	SEPULVEDA AMOR JA	15	2
	SEPULVEDA AMOR J	87	12
	SEPULVEDA J	364	52
HERNÁNDEZ ÁVILA MAURICIO	HERNANDEZ AVILA M	773	118
	HERNANDEZ AVILA MA	45	1
	HERNANDEZ MM	6	2
	HERNANDEZ M	466	83

- *Categoría 3.* Nombres con mayor dificultad de recuperación; mínima precisión. Aquí se incluyeron, por ejemplo, las investigadoras que han cambiado sus apellidos de solteras por el de casadas; o bien, se presentan mezclas de ambos nombres, tal como se describe en la tabla V. En esta categoría es prácticamente imprescindible consultar el currículum vitae de las investigadoras para confirmar la validez de los datos e incrementar la precisión de los resultados.

Tabla V
Categoría 3. Nombres con mayor dificultad de recuperación y mínima precisión en la validación de los datos que se corresponden entre SNI y NCR

Apellido y nombre en el SNI-III	*Datos en NCR*	*Citas*	*Artículos*
OSTROSKY SHEJET MARTHA PATRICIA	OSTROSKY WEGMAN P	701	71
PASANTES ORDÓÑEZ HERMINIA	PASANTES MORALES H	977	79
SCHUNEMANN HOFER ALINE	DEALUJA AS	63	24
VIVIER JEGOUX ANA MARÍA FRANCISCA	VIVIER BUNGE A	88	19

- *Categoría 4.* Nombres del SNI que no coinciden con variante alguna en NCR. Cero recuperación, cero precisión. Para confirmar estos resultados, además de consultar el currículum vitae de los investigadores, es necesario comunicarse con el investigador personalmente para concluir sobre los resultados.

La distribución del número de investigadores que integraron cada una de las categorías arriba descritas se ilustra en la tabla VI. La gran mayoría de los investigadores se agregaron en las categorías 1 (48,33%) y 2 (44,07%). La figura 1 ilustra los flujos de acción tomados para incrementar la precisión y confiabilidad en la validación de datos del SNI, en NCR.

Tabla VI
Distribución del número de investigadores del SNI que fueron encontrados en NCR, para cada una de las categorías establecidas

Categorías	*Núm. de miembros del SNI, en NCR*	*%*
1 Mínima dificultad de recuperación, mayor precisión	318	48,33
2 Mediana dificultad de recuperación, regular precisión	290	44,07
3 Mayor dificultad de recuperación, mínima precisión	4	00,61
4 Cero recuperación, cero precisión	46	06,99
Total	**658**	**100**

Figura 1
Flujos de acción para incrementar la precisión y confiabilidad en la validación de datos del SNI, en NCR

flujo de mayor a menor precisión en la recuperación de datos

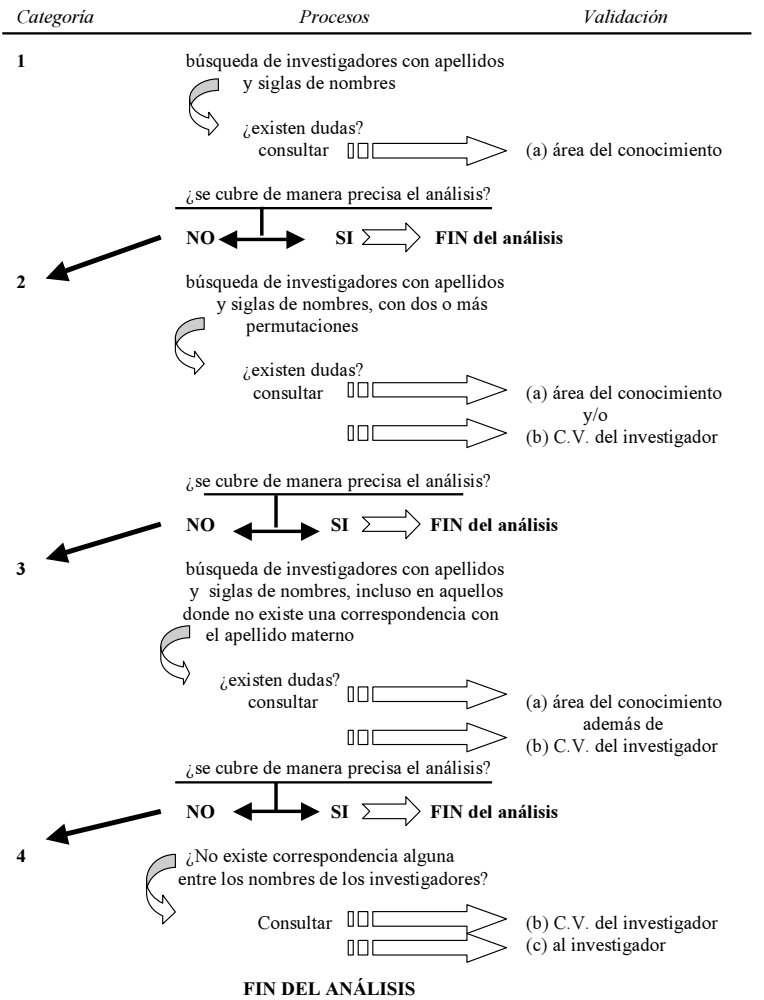
\+

Categoría	*Validación A Por área del conocimiento*	*Validación B C.V. del investigador*	*Validación B consulta al investigador*
1 Mínima dificultad de recuperación, mayor precisión	X		
2 Mediana dificultad de recuperación, regular precisión	X	X	
3 Mayor dificultad de recuperación, mínima precisión	X	X	
4 Cero recuperación, cero precisión		X	X

– *flujo de menor a mayor complejidad en la validación de datos*

Un resultado importante del estudio fue la construcción metodológica de un modelo a seguir para aumentar la recuperación y precisión de los datos en la cobertura de nombres del SNI en NCR. La figura 2 describe el modelo conceptual obtenido. En esta figura se observan tres columnas que definen (1) las cuatro categorías creadas; (2) los procesos a seguir en cada una; y (3) la acción que se puede tomar para validar los datos en cada categoría. En el modelo se puede observar que el análisis de los registros puede terminar en cualquiera de las cuatro categorías. Sin embargo, si se desea ser más exhaustivo en la recuperación y precisión de datos, se tiene la opción de continuar con el resto de las categorías hasta agotar el universo de datos.

Figura 2
Modelo conceptual sobre el uso de la metodología



5. Discusión

En una base de datos, el nombre del autor es un campo importante, utilizado como base de búsqueda y recuperación de información bibliográfica. No obstante, muchos estudios han publicado la falta de uniformidad en los nombres de autores en bases de datos bibliográficas. Al analizar los resultados de estos estudios, se encontró que únicamente han aportado recomendaciones para realizar mejores búsquedas bibliográficas (Kotiaho y cols., 1999; D'Auria, 1997; Shore, 1997; Corrochano, 1996; Sellick, 1996; Meneghini, 1995; Piternick, 1992; Snow, 1986; Piternick, 1985; Pilachowski y Everett, 1985). La evaluación de la calidad de las referencias en los artículos científicos ha revelado que los nombres de los autores son una de las fuentes de errores mayores (Silva, 1992; Sweetland, 1989). De acuerdo a Munley y cols. (1992), una de las razones de las variaciones de nombres personales en bases de datos bibliográficas está relacionada con la diversidad de estructuras que derivan de tradiciones culturales e históricas, involucradas en la forma en que se asignan los nombres en diferentes países. A pesar de la existencia de normas y reglas de catalogación específicas para el manejo de nombres de autores (IFLA, 2002; AACR 1998; RCE, 1995), éstas no son manejadas adecuadamente por la mayoría de las bases de datos bibliográficas. Ruiz-Pérez y cols. (2002), en un estudio exhaustivo, analizaron los orígenes y consecuencias de los errores encontrados en los nombres hispanos, en bases de datos nacionales e internacionales. Sin embargo, sus recomendaciones se limitaron a depurar los criterios de precisión de las bases de datos y a una mejor práctica, por parte de autores y editores de publicaciones, para optimizar la consistencia y confiabilidad en la descripción de nombres. Algunas de estas recomendaciones han sido adoptadas por MEDLINE para el manejo de nombres castellanos (Esteve-Fernández, 2003); sin embargo, Ruiz-Pérez y cols. (2002) concluyen sobre la dificultad del cambio de reglas de indizado de ISI, para acomodar las convenciones lingüísticas del castellano o de otros idiomas diferentes al inglés.

La conducción de este estudio en dos fases ayudó, primero, a identificar la correspondencia general entre los registros del SNI y de NCR, para después seleccionar y analizar una muestra específica de datos con un enfoque exhaustivo. En el proceso de análisis se logró construir una metodología para la identificación de registros de los investigadores del SNI de México en NCR, a través del manejo agrupado de datos. En este proceso se identificaron diversas dificultades relacionadas con la correspondencia de nombres en ambas bases de datos. Por ejemplo, la permutación de hasta siete entradas diferentes para identificar los nombres de los investigadores en NCR (Anexo 1). Este hallazgo obligó a detectar, en paralelo, mecanismos de validación de los registros para incrementar la confiabilidad de los resultados.

Los resultados del proceso de búsqueda, recuperación y validación de nombres de investigadores tienen implicaciones importantes en la evaluación del desempeño de personas, grupos y hasta países. Errores por omisión o confusión de nombres, por

ejemplo, pueden redituar en *sub* o *supravaloración* de la producción, las citas recibidas, y el impacto de la producción de los investigadores. Esto es particularmente relevante para investigadores con nombres hispanos o latinos, donde se utilizan los apellidos paterno y materno, y donde los nombres de pila pueden ser dos o más.

Al establecer una comparación de los resultados obtenidos sobre los miembros del SNI-Nivel III, localizados en NCR, entre las *fases 1 y 2,* se identificó un incremento en la recuperación de registros que fluctuó del 66,11% en la fase 1 del estudio (tabla I), al 93,01% en la *fase 2.* Esta última cifra corresponde a la suma de los registros de las categorías 1 al 3 tal como se describe en la tabla VI. Consideramos que este incremento de 26,9% en la recuperación de registros se debió a la utilización del enfoque creado para la *fase 2* del estudio, y que llevo a la construcción del modelo metodológico descrito. Este hallazgo tiene las dos implicaciones siguientes en el manejo de los registros: (1) el uso de la metodología adecuada redunda en la identificación de un mayor número de registros en NCR; y (2) con el uso de la metodología, aumenta la certeza en la confiabilidad de los resultados.

Si bien es cierto que la muestra de 658 registros correspondientes al nivel III del SNI es pequeña (7,15% del total), la metodología utilizada permite manejar volúmenes mayores de registros a través del agrupamiento de nombres en las categorías señaladas, cuidando además la validación de los registros. Esto es particularmente importante cuando no se tiene acceso inmediato a la currícula de los investigadores, o no se tiene acceso personal al investigador para validar los resultados. De acuerdo a la metodología, sólo en la categoría 4, se debe consultar directamente con el investigador la validación de resultados. A pesar de haber probado el funcionamiento de la metodología en este estudio, se debe continuar explorando su utilidad en el análisis, por ejemplo, de la correspondencia de registros de los otros tres niveles del SNI, en el NCR.

Como resultado de la *fase 2* del estudio, sólo 46 registros del SNI-Nivel III no tuvieron correspondencia alguna con los registros de NCR. Esta cantidad representa 6,99% del total de investigadores de ese nivel. Aun cuando no se ha confirmado este hallazgo con los investigadores, se asume que la producción de este grupo de investigadores se encuentra fuera del período de análisis de este estudio (1984-2002) o bien, la producción de estos autores se encuentra en fuentes de información locales, nacionales o regionales, no incluidas en las bases de datos del ISI. Debido a su alto nivel dentro del SNI, estos 46 investigadores representan un interesante grupo a estudiar, para continuar explorando el uso de metodologías en la evaluación del desempeño de la actividad científica.

6. Conclusiones

Las bases de datos del ISI son utilizadas frecuentemente para conducir estudios métricos de la ciencia y evaluar el desempeño de la actividad científica. El manejo

de los nombres de investigadores en estas bases de datos ha sido una de las dificultades con las que se enfrentan bibliotecarios, analistas y científicos para recuperar de una manera precisa la información buscada. A pesar de ser altamente consultadas las bases de datos del ISI, y de identificar la problemática existente para el manejo de los nombres de los investigadores, no se ha publicado en la literatura un enfoque a seguir para resolver esta situación problemática.

La conducción de este estudio llevó a la construcción de una metodología para la identificación de los nombres de los investigadores del SNI de México, en las bases de datos del ISI. Además del incremento en la recuperación de registros, la metodología incorporó elementos de validación para asegurar la precisión y certeza de los resultados.

Se concluye que la metodología que emergió de este estudio funcionó para lograr el propósito planteado. La exploración sobre el uso de la metodología en el manejo de cantidades mayores de registros ayudará, sin duda, a realizar análisis finales sobre su efectividad en el área.

Finalmente, se reconoce la necesidad de conducir un estudio específico sobre la no correspondencia de 46 investigadores del nivel III del SNI, en NCR.

7. Referencias

AACR. *Anglo-American Cataloguing Rules*, 2.ª ed. Ottawa: Canadian Library Association. London, U.K.: Library Association Publishing. Chicago: American Library Association, 1998.

AGUILLO, I. F. Cybermetrics/Webometrics: an emerging discipline. En: *VIII ISSI (International Society for Scientometrics and Informetrics) Conference*. Sidney, Australia. 16-20 de julio de 2001.

ALMEIDA, N. (2003). Research on health inequalities in Latin America and the Caribbean: bibliometric analysis (1971-2000). *Am J Public Health,* vol. 93 (12), 2037.

BIREME. Centro Latinoamericano y del Caribe de Información en Ciencias de la Salud, 2004. Recuperado el 13 de agosto de 2003 en: http://www.bireme.br/bvs/E/ehome.htm

BORDONS, M.; ZULUETA, M. A. (1999). Evaluación de la actividad científica a través de indicadores bibliométricos. *Revista Española de Cardiología,* vol. 52, 790-800.

CASTRO, R.; CASTRO, F.; MUGNANIAI, R. Afiliación de autores y títulos de revistas en los estudios bibliométricos desde las bases de datos MEDLINE, LILACS y SciELO. En: *II Seminario internacional sobre estudios cuantitativos y cualitativos de la ciencia y la tecnología «Prof. Gilberto Sotolongo Aguilar» INFO 2004*. La Habana, Cuba.

CETTO, A. M.; ALONSO-GAMBOA, O. (1998). Scientific Periodicals in Latin America an the Caribbean: A Global Prespective. *Interciencia,* vol. 23 (2), 84-93.

COLLAZO-REYES, F.; LUNA-MORALES, M. E. (2002). Física mexicana de partículas elementales: organización, producción científica y crecimiento. *INCI,* 27 (7), 347-353.

CORROCHANO, L. M. 1996. Spanish practice. *Nature,* vol. 384 (6605), 106.

COSTAS-COMESAÑA, R.; GARCÍA-ZORITA, J. C. Indicadores de rendimiento en las

bases de datos bibliográficas: la tasa de filtrado del campo de autor. Una aplicación al caso de nombres de autores españoles. En: *II Jornadas de Tratamiento y Recuperación de la Información (JOTRI).* Universidad Carlos III de Madrid. 8 y 9 de septiembre 2003.

COSTAS, R.; BORDONS, M. Methodological procedure to overcome the lack of normalisation of author names in bibliometric analyses at the micro level. *Proceedings of ISSI 2005: the 10th International Conference of International Society for Scientometrics and Informetrics. Stockolm:* Karolinska University Press, 2005, p. 688.

CRONIN, B. *The Citation process. The role and significance of citation in scientific communication.* London: Taylor Graham, 1984.

D'AURIA, D. (1997). Six characters in search of an author (ed.). *Occup Med* (Oxford), vol. 47 (4), 195.

DIARIO OFICIAL DE LA FEDERACIÓN. Acuerdo de Creación del Sistema Nacional de Investigadores. México, 26 de julio de 1984. DOF, 9.

DRENTH, J. P. H. (1998). Multiple authorship. The contribution of Senior Authors. *JAMA,* 280, 219-221.

EPSTEIN, R. J. (1993). Six authors in search of a citation: villains or victims of the Vancouver convention? *BMJ,* 306, 765-767.

ESTEVE FERNÁNDEZ, A. M. G. (2003). Accuracy of referencing of Spanish names in Medline. *Lancet,* 361, 351-352.

EUROPEAN COMMISSION. Third European Report on S&T Indicators: Towards a Knowledge-based Economy. Brussels: European Commission, 2003.

FLANAGIN, A.; CAREY, L. A.; FONTANAROSA, P. B.; PHILIPS, S. G.; PACE, B. P.; LUNDBERG, G. D.; RENNIE, D. (1998). Prevalence of articles with Honorary Authors and Ghost Authors in peer-reviewed medical journals. *JAMA,* vol. 280, 222-224.

GARFIELD, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics,* 1, 359-375.

GARFIELD, E. (1990). How ISI selects journals coverage: quantitative and qualitative considerations. *Current Contents,* 22, 5-13.

GARFIELD, E.; WELLJAMS-DOROF, A. (1992). Citation data: their use as quantitative indicators for science and technology evaluation and policy-making. *Science and Public Policy,* 19, 321-327.

GÓMEZ, I.; BORDONS, M. (1996). Limitaciones en el uso de los indicadores bibliométricos para la evaluación científica. *Pol Cient,* 46, 21-26.

GÓMEZ, I.; COMA, L.; MORILLO, F.; CAMI, J. (1997). Medicina Clínica (1992-1993) vista a través del Science Citation Index. *Med Clin (Barc),* vol. 109 (13), 497-505.

HAMILTON, D. P. (1991). Research papers: who's uncited now? *Science,* 251, 25.

IFLA. *Names of persons: national usages for entries in catalogues.* London: IFLA International Office for UBC, 2002, pp. 39-41.

ISI. National Citation Report. [Recuperado el 13 de agosto de 2003 en: http://www. isinet.com//isi/producst/rsg/products/ncr/index.html.]

KOTIAHO, J. S.; TOMKINS, J. L.; SIMMONS, L. W. (1999). Unfamiliar citations breed mistakes. *Nature,* vol. 400 (6742), 307.

LICEA, J.; SANTILLÁN-RIVERO, E. (2002). Bibliometría ¿para qué? *Bibli Univ,* vol. 5 (1), 3-10.

LÓPEZ-CÓZAR, E. (1997). Incidencia de la normalización de las revistas científicas en la

transferencia y evaluación de la información científica. *Rev Neurol,* vol. 25 (148), 1942-1946.

MACROBERTS, M. H.; MACROBERTS, B. R. (1996). Problems of citation analysis. *Scientometrics,* 36, 435-444.

MACÍAS-CHAPULA, C. A. (1991). Análisis de citas de cuatro revistas biomédicas latinoamericanas. *Revista Española de Documentación Científica,* vol. 14 (4), 220-227.

MACÍAS-CHAPULA, C. A. (1994). Non-SCI subject visibility of the Latin America scientific production in the health field. *Scientometrics,* 30 (1), 97-104.

MACÍAS-CHAPULA, C. A. (1995). Primary heath care in Mexico: a «non- ISI» bibliometrics analysis. *Scientometrics,* vol. 34 (1), 63-71.

MACÍAS-CHAPULA, C. A.; RODEA-CASTRO, I. P. (1997). Subject content of the Mexican production on health and the environment. *Scientiometrics,* vol. 38 (2), 295-308.

MACÍAS-CHAPULA, C. A.; RODEA-CASTRO, I. P.; NARVÁEZ-BERTHELEMONT, N. (1998). Bibliometric analysis of AIDS literature in Latin American and the Caribbean. *Scientometrics,* vol. 41 (1-2), 41-49.

MACÍAS-CHAPULA, C. A. (2002). Bibliometric and webometric analysis of health system reforms in Latin American and the Caribbean. *Scientometrics,* vol. 53 (3), 407-427.

MENEGHINI, R. (1995). Systematization of academic and scientific affiliation, or how to prevent data on your publications from being lost in the national and international databases. *Braz J Med Biol Res,* 28 (6), 617-619.

MOED, H. F. *The use of bibliometric indicators for the assessment of research performance in the natural and life sciences.* Leiden: DSWO Press, 1989.

MORAVSIK, M. J. (1989). ¿Cómo evaluar a la ciencia y a los científicos? *Revista Española de Documentación Científica,* 12, 313-325.

MUNLEY, P. H.; ANDERSON, M. Z.; BAINES, T. C.; BORGMAN, A. L.; BRIGGS, D.; DOLAN, J. P. jr.; KOYAMA, M. (2002). Personal dimensions of identity and empirical research in APA journals. *Culture Divers Ethnic Minor Psychol,* Nov., vol. 8 (4), 357-65.

NATIONAL SCIENCE BOARD. *Science & Engineering Indicators.* Washington, DC, US Government Printing Office (NSB 96-211), 1996.

OCDE. *Main Science & Technology Indicators.* París: OCDE, 2002.

PELLEGRINI, FILHO A.; GOLDBAUM-M. S. J. (1997). Production of Scientific articles about health in six Latin American contries, 1973-1992. *Rev. Panam Salud Pública,* 1 (1), 23.

PILACHOWSKI, D. M.; EVERETT, D. (1985). What's in a name? Looking for people online-social sciences. *Database,* 8 (3), 47-65.

PITERNICK, A. B. (1985). What's in a name? Use of names and titles in subject searching. *Database,* 8 (4), 22-28.

PITERNICK, A. B. (1992). Name of an author! *Indexer,* 18 (2), 95-100.

RENNIE, D.; YANK, V.; EMANUEL, L. (1997). When authorship fails. A proposal to make contributors accountable. *JAMA,* 278, 579-85.

RCE. *Reglas de Catalogación Españolas.* Madrid: Dirección General del Libro, Archivos y Bibliotecas, 1995, pp. 431-454.

REYES, B. H.; KAUFFMANN, Q. R.; ANDERSEN, H. M. (2000). La autoría en los manuscritos publicados en revistas biomédicas. *Rev Med Chile,* 128 (4), 363-366.

RICYT. *Indicadores Iberoamericanos de Ciencia y Tecnología.* RICYT: Buenos Aires, 2002.

RUIZ-PÉREZ, R.; DELGADO LÓPEZ-CÓZAR, E.; JIMÉNEZ-CONTRERAS, E. (2002). Spanish personal name variations in national and international biomedical databases: implications for information retrieval and bibliometric studies. *J Med Libr Assoc,* 90 (4), 411-430.

RUSSELL, J. M. *Collaboration and research reference in science: a study of scientists at the National University of Mexico (UNAM).* PHD Thesis. Dep. of Information Science, City University London, 1998.

SANCHO, R. (1990). Indicadores bibliométricos utilizados en la evaluación de la ciencia y la tecnología. Revisión bibliográfica. *Revista Española de Documentación Científica,* 13 (3-4), 842-865.

SCOVILLE, C. L.; JOHNSON, E. D.; MC CONNELL, A. L. (2003). When A. Rose is not A. Rose: the vagaries of author searching. *Med Ref Serv Q,* 22 (4), 1-11.

SELLICK, J. T. C. (1996). Multiple authors. *Nature,* 383 (6601), 569.

SHORE, M. L. (1997). Variation between personal name headings and title page usage. *Cat Class Quart,* 4 (4), 1-11.

SIEBERS, R.; HOLT, S. (2000). Accuracy of references in five leading biomedical journals. *Lancet,* 356, 1445.

SILVA, G. A. (1992). Nombres de pila completos: las iniciales no bastan. *Med Clin* (Barc.), 99 (11), 435.

SMALL, H. (1999). Visualizing Science by Citation Mapping. *J Am Soc Information Science,* 50 (9), 799.

SMALL, H. (2003). Paradigms, Citations, and Maps of Science: A Personal History. *J Am Soc Information Science,* 54 (5), 394.

SNOW, B. (1986). Caduceus: people in medicine names online. *Online,* 10 (5), 122-127.

SWEETLAND, J. H. (1989). Errors in bibliographic citations: a continuing problem. *Libr Quart,* 59 (4), 291-304.

TORVIK, V. I.; WEEBER, M.; SWANSON, D. R.; SMALHEISER, N. R. (2005). A probabilistic similarity metric for Medline records: a model for author name disambiguation. *J Am So Inf Sc & Tech,* vol. 56 (2), 140-158.

VAN RAAN, A. F. J. (1993). Advanced bibliometric methods to assess research performance and scientific development: basic principles and recent practical applications. *Research Evaluation,* 3, 151-166.

VELHO, L. (1990). Indicadores científicos: en busca de una teoría. *Interciencia,* 15(3), 139-145.

WOODING, S.; WILCOX-JAY, K.; LEWISON, G.; GRANT, J. (2004). Co-Author Inclusion: a novel recursive algorithmic method for dealing with homonyms in bibliometric analysis. Eighth International Conference on Science and Technology Indicators. Book of abstracts program. Leiden: CWTS, Leiden University

WILCOX, L. J. (1998). Authorship: the coin of realm, the source of complaints. *JAMA,* 280, 216.

ZULUETA, M. A.; BORDONS, M. (1999). La producción científica española en el área cardiovascular a través de Science Citation Index (1990-1996). *Rev Esp Cardiol,* 52, 751-764.

ANEXO 1

Relación de variantes o permutaciones encontradas a los nombres de los investigadores en (1) la base de datos del Sistema Nacional de Investigadores (SNI) de México, y (2) el *National Citation Report* (NCR), del *Institute for Scientific Information* (ISI).

Variantes en la base de datos del SNI:

1. Apellido paterno, apellido materno y uno o más nombres de pila.

Variantes en la base de datos del NCR:

1. Apellido paterno, apellido materno y una sigla del nombre de pila.
2. Apellido paterno, apellido materno y dos siglas de los nombres de pila.
3. Apellido paterno y una sigla del nombre de pila.
4. Apellido paterno y dos siglas de los nombres de pila.
5. Apellido paterno y una de las dos siglas de los nombres de pila.
6. Apellido paterno, una sigla del nombre de pila y una sigla del apellido materno.
7. Apellido paterno, una sigla del nombre de pila, una sigla del apellido materno y una sigla desconocida o no reconocida por el SNI.